Trusting Crowdsourced Information on Cultural Artefacts

Nottamkandath, A.

2016

document version
Publisher's PDF, also known as Version of record

Archana Nottamkandath

Trusting Crowdsourced Information on Cultural Artefacts

SIKS Dissertation Series No. 2016-09

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

VRIJE UNIVERSITEIT

Trusting Crowdsourced Information on Cultural Artefacts

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op dinsdag 29 maart 2016 om 13.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door
Archana Nottamkandath
geboren te Palghat, India

promotor: prof.dr. W.J. Fokkink

Promotiecommissie:

Dr. L.M. Aroyo

Prof.dr. G.J. Houben

Dr. A. Isaac

Dr. J. Leidner

The truth is, most of us discover
where we are heading when we arrive.

Bill Watterson

PREFACE

"Isn't it funny how day by day nothing is different, but when you look back everything is different". The entire PhD journey was such an experience.

I would like to thank my advisor Wan Fokkink for being constantly available, providing strong direction, offering invaluable feedback and always encouraging me throughout the PhD. Although the overall goal of this thesis was known from the beginning, Wan has given me the freedom to explore different areas of interest while emphasizing the importance of quality and research. This thesis is a result of such explorations through many pathways while always having the horizon in sight.

Davide was always helpful and a source of inspiration since the beginning of my PhD. I very much value the many stimulating discussions we had on the topic of trust as well as our successful collaborations. Jasper has been a well-organised and motivating colleague and our collaborations helped in exploring different techniques for solving issues related to trust. The dataset from an experiment performed by CWI team comprising Jacco, Myriam and Mieke was used for evaluations in one of the chapters. Gerben de Vries from University van Amsterdam was a co-author for the chapter where we used RDF graph kernels. Paul and Willem have also been my co-authors. I really enjoyed working with you all. The discussions and experiences from these collaborations are valuable lessons learnt.

It has been a great pleasure working with the COMMIT SEALINCMedia project team which was a collaboration between universities and cultural heritage institutions. I would like to thank all the members of the project and Lizzy and Hendrike from the Rijksmuseum. Collaborations with SEALINCMedia PhD students and their advisors helped in putting together different pieces of the bigger puzzle we were all working on and helped attain a bigger perspective on problem solving. Thanks to the entire team: Chris, Jasper, Mieke, Myriam, Jacco, Lora, Geert-Jan, Antoine, Alessandro, Guus and Jan. Apologies if I have missed out anyone.

It was a pleasure to have Chris and Valentina as my office mates. Chris, I cannot thank you enough for all the times you have helped me "integrate" and for the translation of letters from every possible organization in The Netherlands. Laurens, it was nice to have you constantly innovate my workplace and occasionally demonstrate how to keep my whiteboard "white". Jokes apart, thank you for always helping me with different

tasks and for providing valuable suggestions. Cristina, you have been a really sweet and encouraging friend and it was really nice to have you at our office. The Semantic Web group floor is where I spent my time and they are a bunch of active and fun colleagues.

During my internship at Thomson Reuters in London, I had the pleasure to work with Jochen Leidner and his team: Vassilis, Chris, Tim and Ola. It was a wonderful experience to learn in a new environment and special thanks to Jochen for making this happen.

A special thanks to my reading committee Lora Aroyo, Geert-Jan Houben, Antoine Isaac and Jochen Leidner for their efforts to read and accept my thesis and for providing invaluable feedback.

My parents and brother have always supported during my studies and their motivation and encouragement has helped me come a long way. My aunts Viji and Shanthi (a special thanks to you both!), uncles, grandparents and cousins have always encouraged me in my ventures and a big thanks to them for all this!

Life in Amsterdam has been a joy mainly because of my friends. Ahmad, Qaisar and Bilal, it is wonderful to have you guys around. Fouad, it was always fun to meet you during your Amsterdam trips. Melani, I am very glad to have met you during my early days in Amsterdam and I was very happy that I could spend time with you again last year. Indu, I had a great time meeting you during the different trips, come again soon!

This thesis is not an end to a journey, but is a fulfilling chapter among the many others which have been and those that are yet to begin. One of the best lessons I have learnt is to follow your heart and to work towards keeping your passion alive. Thus signing off with a motivating quote, *"Stay hungry, stay foolish".*

Contents

Chapter 1

Introduction

1.1 Motivation

Cultural and heritage preserving organisations such as museums are rapidly digitising their collections, and at the same time migrating digitised collections to the Web. Through the Web, these institutions can reach large masses of people, with intentions varying from increasing visibility (and hence visitors) to acquiring user-generated content.

The physical artefacts preserved by the institutions have some basic information such as title, dimensions, information about materials and techniques used to create the artefact etc.. This list of information about the artefacts vary in their completeness and various institutions gather and store different aspects of information which they deem relevant to physically store and retrieve the artefacts in their collection. The digitisation process involves photographing the artefacts and storing their relevant information in an online system of the cultural heritage institutions. Digitisations has also helped manage the limitations of physical space, conservation, location and opening hours that previously affected access to collections [85]. Cultural institutions use different models such as the Digital Content Life Cycle Model[1] which encapsulates the main activities carried out by cultural heritage organisations, from selecting to creating, managing, discovering, using and reusing (including licensing) as well as preservation. In order to facilitate archiving and retrieval operations on the Web, collections must be described by high-quality information that cover physical properties (e.g. dimensions, material), provenance (e.g. creator, previous owners) and subject matter (e.g. what is represented) of the artworks. The process is called annotating and the information provided is called an annotation. The annotations later become the metadata for the artworks in the institutions and

[1] `http://www.digitalnz.org/make-it-digital`, January 2016

help various users, both internal and external to the museums, search for artefacts online. Annotations can be either in the form of specific tags or free text and describe the artefacts.

During the digitisation process, cultural heritage institutions employ professionals, mostly art historians, to provide high-quality annotations about artefacts. They are trained and follow strict guidelines on how to correctly and qualitatively annotate artefacts; but their effectiveness is hindered by different factors such as the size of museum collections (which can be in the order of millions of artworks), time or monetary constraints of cultural institutions, and lack of domain expertise on some of the subject matter of artworks.

Crowdsourcing, originally described as the act of taking work once performed within an organisation and outsourcing it to the general public through an open call for participants[2], is becoming increasingly common in museums, libraries, archives and the humanities as a tool for digitising or computing vast amounts of data [86]. Crowdsourcing of cultural heritage collections has been described extensively in the work of Mia Ridge [86]. In this thesis we will refer to excerpts from their work for providing better detail about the motivation of crowdsourcing in the cultural heritage domain. In that book it is mentioned, "Crowdsourcing in cultural heritage benefits from its ability to draw upon the notion of the "greater good" in invitations to participate, and this may explain why projects generally follow collaborative and co-operative, rather than competitive, models. Crowdsourcing in cultural heritage is more than a framework for creating content: as a form of engagement with the collections and research of memory institutions, it benefits both audiences and institutions."

Carletti et al. explored digital humanities and crowdsourcing in their work "Digital Humanities and Crowdsourcing: An exploration"[3]. The various challenges and opportunities for crowdsourcing in the cultural heritage domain are listed out in detail by Oomen and Aroyo [78]. Crowdsourcing initiatives have been classified into categories and the technique referred to in this thesis of gathering descriptive metadata related to objects in a collection is considered of the crowdsourcing type called as "classification". However, issues have been raised about the quality of crowdsourced contributions. Oomen and Aroyo mention "Cultural heritage institutions earned their reputation over the years by preserving the quality and truthfulness of the information they offered by having full control over the acquisition, organization and the annotation of the collection items." Having information contributed from the crowd could be seen as a threat to their authoritative position. Some examples showing that cultural heritage institutions are concerned

[2]Howe, J. The Rise of Crowdsourcing; `http://www.wired.com/2006/06/crowds/`, January 2016

[3]`http://mw2013.museumsandtheweb.com/paper/digital-humanities-and-crowdsourcing-an-exploration-4/`, January 2016

about the quality of information are the following. Nineteenth-century natural historians corresponding with amateur observers about the distribution of botanical specimens had to try to determine the veracity and credibility of their contributions [90]. Modern manuscript translation projects such as *Transcribe Bentham*[4] initially questioned the editorial quality of volunteer-produced transcripts.

Cultural heritage institutions need the annotations to be trustworthy in order to maintain their authoritative reputation. Various case studies have shown that there is a wide diversity in the type of information provided on the Web and also in its quality. For instance, the artwork collection item (a sculpture) from a the `Steve.Museum` dataset[5] in Figure 1.1 is depicted with the annotations produced by crowd annotators in a real-world annotation campaign. The annotations in green indicate ones which were considered useful by professionals at institution, while the red ones indicate ones which were not considered useful. From the figure we can see that there is variability about usefulness of the same annotation by professionals in the same institution. The annotation "gold" was considered useful by one professional while another professional did not find it useful. The disagreement may have arisen since one reviewer would have considered any annotation relevant to the artwork as relevant, while the other one would have expected a more specific annotation for the artwork. This issue of variability in opinions of usefulness by professionals can be overcome by having more detailed guidelines for reviewing annotations or by having techniques to resolve this issue once it arises. This shows that evaluation of annotations from the crowd is a challenging task. Employing human reviewers to assess the quality of annotations is as expensive as hiring professional annotators and requires additional verification of their skills to function as reviewers.

FIGURE 1.1: The artwork titled *Kinarra* from the `Steve.Museum` dataset and associated crowd annotations. Green represents useful annotations while red represents non-useful annotations.

[4] `http://www.ucl.ac.uk/Bentham-Project`, January 2016

[5] `http://www.steve.museum/`; last accessed on 1st November 2011 for downloading dataset. The dataset is no longer available to be downloaded from the listed domain. However, the data used in our experiments has been made available online.

Digital technologies help in facilitating data gathering and feedback, computationally validate contributions and provide ability to reach both broad and niche groups. Challenges that are related to the quality of data was explored by Oomen and Aroyo [78] which are: maintaining/resolving conflicting information, maintaining and presenting extensive provenance information, creating open and clear reviewing procedures, evenly distributing the contributions of the users over the entire collection and indicating when an annotation is "good" or "finished". In our work we do not deal with techniques that aim at identifying the best potential users from the crowd, techniques to understand user incentives or that aim at designing effective annotation tasks. We focus on the information that is collected and we observe how the cultural heritage institutions perceive that information from a qualitative perspective. This observation is then utilised to develop algorithms which can (semi-)automatically help determine quality of annotations and annotators over the Web.

The general definition of trust that is used in this thesis is from Olmedilla et al. [77] and is defined as:

"Trust of a party A to a party B for a service X is the measurable belief of A in that B behaves dependably for a specified period within a specified context (in relation to service X)."

According to this definition, trust is evaluated relative to a specific service. In our case, party A refers to employees of the cultural heritage institutions and party B refers to annotators from the Web. Service X refers to the annotation process. Trust is a subjective phenomenon and humans use the concept of trust in various situations. Recommendations about movies and books from certain friends are valued more than others. In many cases, trust of data is linked to the reputation of the source. People are more likely to trust a certain person if they had a positive experience in the past. Artz et al. [4] define reputation as:

"Reputation is an assessment based on the history of interactions with or observations of an entity, either directly with the evaluator or as reported by others (recommendations or third party verification)."

Many systems, especially on the Web, choose to reduce trust to reputation estimation and analysis alone. However, trust can also be based on many other factors such as a real-world experience with the source, beliefs based on stereotypes of the source (such as age, gender etc.), knowledge of how the data was produced (provenance), guarantee from a trusted third-party. In this thesis we initially model trust based on reputation of annotators and then proceed to explore modelling and computation of trust based on some of the enlisted factors.

Once the cultural heritage institutions obtain the annotations from annotators on the Web, they assign a level of quality to the annotation. The most widely used definition of quality is "fitness of use". However, Gamble et al. [34] argue that this leads to assertion that quality of data cannot be assessed independent of consumer. In the case of cultural heritage domain, the consumers are the users (both internal to the organisations and on the Web) who use the annotations to search and retrieve artefacts. However, there would be annotations which are relevant to the artefact but which are not used by the crowd to search for that artefact. Thus we use a more objective definition of quality as defined in ISO 9001, which is:

"The degree to which a set of inherent characteristics fulfil requirements".

The levels of quality is subjective for different institutions. In some cases it can be as simple as "good" and "bad", while in others there might be more detailed classification such as "useful", "typo", "foreign-language", "judgement", "not-useful", etc.. The level of quality assigned to an annotation is defined by the policies of that institution. Artz et al. [4] define policies as:

" Policies describe the conditions necessary to obtain trust, and can also prescribe actions and outcomes if certain conditions are met."

Thus cultural heritage institutions use policies to decide whether or not to trust an annotator for the annotation task and also to determine the quality level of annotations. Policies can be very detailed and can mention further actions such as deciding what status to give to the annotator based on his/her performance, whether or not to add annotations to their collections, etc..

In this thesis, our challenge is to build algorithms which help to predict quality of annotations and reputation of annotators. In the next section we provide the context to our work along with background of crowdsourcing and trust in cultural heritage context.

1.2 Background

This section presents the context of the research presented in this thesis, as well as a general discussion of the background and related work.

1.2.1 Context

The work in this thesis is part of Socially Enriched Access to Linked Cultural Media (SEALINCMedia), a sub-project of the COMMIT project in The Netherlands which is

a collaboration between academic institutions and industry. As part of this project, Rijksmuseum[6] in Amsterdam aims to enrich their collection by obtaining annotations through crowdsourcing. This would enable better search and retrieval in their online system. The goal of the project is to create a platform to facilitate the annotation process and involves various components such as finding experts on the Web to annotate their collection, recommending the artworks for annotation, providing an annotation interface, evaluating the quality of provided annotations, and tracking performance of annotators. The platform *Accurator*[7] was developed as part of the project to facilitate crowdsourcing of the Rijksmuseum collection.

1.2.2 Crowdsourcing annotations of cultural artefacts

Crowdsourcing techniques are widely used by cultural heritage and multimedia institutions for enhancing the available information about their collections. Examples include the Tag Your Paintings project [29], the `Steve.Museum` project[8], the *Waisda?* video tagging platform[9], ESP game [100], and others such as Brooklyn Museum and the New York Library [86]. The Smithsonian Institution also has a long history with 'proto-crowdsourcing'[10].

The success of these initiatives clearly shows the potential of crowdsourcing techniques for artefact annotation purposes, but also highlights many challenges. Different from professional annotators, crowd annotators have very limited annotation guidelines. They also lack tertiary education in art history or lack professional experience in cultural heritage curation. Apart from that, their background, skills and expertise are not known in advance; therefore, their performance in the system cannot be guaranteed in a straightforward manner. Mechanisms need to be designed to determine trust in crowdsourced annotations.

1.2.3 Crowdsourcing and trust

Studies have been done to understand the quality of information provided by the crowd, as shown by Snow et al. [94] and Aroyo et al. [3]. Inel et al. [50] studied the annotations obtained from crowdsourcing platforms such as Crowdflower to make quality assessments. Many methods have been developed to determine the quality of such crowdsourced information, where majority voting has been widely used. For example, in the ESP game,

[6] `https://www.rijksmuseum.nl`, January 2016
[7] `http://annotate.accurator.nl/`
[8] `http://www.steve.museum/`; last accessed on 1st November 2011 for downloading dataset
[9] `http://waisda.nl`
[10] `http://siarchives.si.edu/blog/smithsonian-crowdsourcing-1849`

a label is added to the picture if at least two randomly picked users suggest the same label.

Trust management in crowdsourced systems often employs wisdom of the crowd approaches [96]. Wisdom of the crowd is the collective opinion of a group of individuals rather than of a single expert. However in our scenario, this approach will not be effective since annotations for an artefact by different annotators can be very diverse due to difference in background and knowledge of annotators . General consensus techniques work better when the annotations are of a more general nature. Gamification, which is the application of game-design elements and game principles in non-game contexts, is another approach that leads to an improvement of the quality of annotations gathered from crowds, as shown, for instance, in von Ahn et al. [100]. The work presented here is orthogonal to a gamified environment, as it allows us to semi-automatically evaluate the user-contributed annotations and hence to semi-automatically incentivise them. By combining the two, museums could increase the user incentivisation (showing his reputation may be enough to incentivise a user) while curating the quality of annotations. Users that participated in the experiments that provided the datasets for our analyses did not receive monetary incentives, so leveraging incentives related to gamification and personal satisfaction (by means of reputation tracking) may reveal to be an important factor in increasing the accuracy of the annotations collected.

In folksonomy systems such as the `Steve.Museum` project, tag evaluation techniques such as comparing the presence of the annotations in standard vocabularies and thesauri, determining their frequency and their popularity or agreement with other annotations (see, for instance, Van Damme et al. [24]) have been employed to determine the quality of annotations entered by annotators. Such mechanisms focus mainly on the contributed content with little or no reference to the user who authored it. Also, in folksonomy systems the crowd often manages the annotations, while in our scenarios the crowd only provides the annotations, which are managed by museums or other institutions, according to specific policies. Medeylan et al. [71] present algorithms to determine the quality of annotations entered by annotators in a collaboratively created folksonomy, and apply them to the dataset CiteULike[11], which is a website for bibliographic references with user-annotated folksonomy tags. They evaluate the relevance of user-provided annotations by means of text document-based metrics. In our work, since we evaluate annotations, we cannot apply document-based metrics, and since we do not have at our disposal large amounts of annotations per subject, we cannot check for consistency among annotators tagging the same image. We do not have well trained image analysis software or explicit museum policies, so it is hard to distinguish if possible conflicts between annotations regarding the same image are due to the fact that some are correct and some not, or

[11] `http://www.citeulike.org/`

that they refer to different aspects (or parts) of a complex picture. Therefore, instead of assuming one of the two cases a priori, we determine the trustworthiness of annotations on the basis of the reputation of their user or provenance stereotypes. Provenance-based techniques have been used by Ceolin et al. [16] to determine trust of event descriptions. In open collaborative sites such as Wikipedia[12], where information is contributed by Web users, automated quality evaluation mechanisms have been investigated (see, for instance, De La Calzada et al. [25]). Most of these mechanisms involve computing trust from article revision histories and user groups (see Zeng et al. [109] and Wang et al. [102]). These algorithms track the changes that a particular article or piece of text has undergone over time, along with details of the annotators performing the changes.

Majority voting [47] is a commonly used method to assess the quality of annotations. But it has proven only partially effective in the cultural heritage scenario, mainly due to the sparseness issue. It is difficult to obtain multiple same annotations for an artwork due to the diversity in knowledge and background of annotators as discussed earlier [60]. Some items might contain abstract/fictional/factual elements that are hard to be recognised and described without proper knowledge. Also cultural heritage institutions have a wide variety of artefacts and art is a subjective concept to annotators. They provide annotations concerning different aspects of the artefacts such as creator, date of creation, time, place, title, visual representation, factual description, sentiments, relevance of an artefact, details about possession of an artefact, etc., due to which there is little chance for a majority agreement. Adapted annotator agreement or disagreement measures have also been studied [35, 49], by considering, for example, annotator history and agreement with aggregated labels. In *Waisda?* [46] annotations on videos are considered trustworthy if entered by two different annotators within a certain time interval. In the "Your Paintings Tagger" [29] a tag is accepted if it has been employed in annotations of an image ten different times.

1.2.4 Trust and reputation

Trust is a widely explored topic within a variety of computer science areas. Sabater and Sierra [87] have studied about computational trust and reputation models in the area of distributed Artificial Intelligence. They studied different classification dimensions for trust and reputation models. From the various models presented, the work used in this thesis aligns with the trust model proposed by Castelfranchi and Falcone [14], which is a cognitive trust model. In models based on a cognitive approach, trust and reputation are based on underlying beliefs and are a function of the degree of these beliefs.

[12] `http://www.wikipedia.org`

Artz and Gil [4] provided an overview of existing trust research in computer science and the Semantic Web. Trust is a central component of the Semantic Web vision [5–7]. The Semantic Web stack [6] has included all along a trust layer to assimilate the ontology, rules, logic and proof layers. Artz and Gil report that on the Web important trust judgements are in the hands of humans. However, in the Semantic Web, both humans and agents will be the consumers and agents will need to automatically make trust judgements to choose a service or information source while performing a task.

Another interesting work is that of Golbeck and Hendler [37, 38] where they describe an application call *TrustMail* which uses trust decision as a transitive process. This means that trusting one piece of information or source requires trusting another associated source. We use the concept of transitive trust for determining quality of an annotation provided by an annotator on a new topic, say t', based on their performance for an previously provided annotation, say t, which is semantically closest to t'.

Prasad et al. [81] presented a comparative analysis of trust (models and metrics) in diverse contexts and provided a comprehensive ontology to capture trust-related concepts. The also provided details of the theoretical underpinnings and comparative analysis of Bayesian approaches to binary and multi-level trust, to automatically determine trustworthiness in a variety of reputation systems including those used in sensor networks, e-commerce, and collaborative environments. They discussed about Beta probability distributions and Dirichlet distributions and subjective logic techniques which are used in this thesis are also equivalent to beta probability techniques for binomial distributions and to Dirichlet distributions for multinomial distributions as will be explained in detail in Chapter 2.

O'Hara [76] provided an overview of conceptual analysis of trust. He identified the key parameters that enable to investigate and identify trust, thereby enabling to develop systems, institutions and technologies to support, model or mimic trust. The goal of our work is also the same. We aim to determine trust based on evidence gathered over time about the annotators and also by using different features such as information in the user profile and provenance information to build trust models for cultural heritage domain. In his work he emphasises the importance of reputation of users for determining trust which is also an underlying assumption for our techniques and evaluations.

Reputation of annotators can be computed based on their actions in the past. Reputation is an important mechanism in our set of strategies to place trust since it can be used as an additional input to predict quality of annotations; annotators with a high reputation are more likely to create high-quality annotations. In addition, reputation scores can be used by institutions for task allocation. Tasks of higher complexity can be allocated to highly reputed annotators. Annotator reputation can also be used to grant more privileges to

annotators, for example, tasks of reviewing annotations, a technique that is successful on popular platforms such as Stackoverflow[13]. Modelling of reputation and annotator behaviour on the Web is a widely studied domain. Javanmardi et al. [51] propose three computational models for user reputation by extracting detailed user edit patterns and statistics which are particularly tailored for wikis, while we focus on the annotations domain.

Reputation of annotators can be constantly updated based on new evidence obtained. This helps to track annotators and utilise their latest value of reputation to compute quality of annotations that they will provide in the future.

Another approach to obtain trustworthy data is to find experts amongst Web users with a good intention (see De Martini et al. [27]). This mechanism assumes that annotators who are experts tend to provide more trustworthy annotations. It aims at identifying such experts, by analysing the profiles built by tracking annotator performance. Breslin et al. [9] use internet-based discussions to find experts in online communities and associated social networks while Zhou et al. [110] describe techniques to route questions to the right users in online communities and Cosley et al. [23] describe how to use intelligent task routing to help people find work in Wikipedia.

In our work, we have annotations from completed tasks performed by annotators from the Web and we build reputations of these annotators based on evidence available at hand as well as by employing their demographics information.

1.3 Techniques

We describe the various techniques we will be using throughout this thesis to discuss about how we model, represent and predict quality of annotations and reputation of annotators.

1.3.1 Representing trust

In order to determine the quality of annotations, we need to model the various entities involved in the process of producing an annotation such as the annotator, and details of how the annotation was produced such as vocabularies used, timestamp, day of creation, etc. Quality of annotations must be modelled as a quantifiable term from institution policies and standards. In our work we use the Semantic Web for modelling annotations, annotators, the annotation process and quality of annotations and annotators. “The

[13] `http://www.stackoverflow.com/`

Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF)."[14]

Ontologies are available for the Semantic Web to classify the terms that can be used in a particular application, characterise possible relationships, and define possible constraints on using those terms. We use the Open Annotation (OA) ontology [88] to model annotations and its properties. It models the annotator, the artefact for which the annotation was produced and the annotation term itself. The quality of an annotation is modelled as an annotation of an annotation with the quality score as the annotation term.

The Provenance Ontology (PROV-O)[15] helps to records different aspects of the annotation creation process such as timestamp of creation, typing duration, annotator, activity for which the annotation was produced, etc.

In order to model the annotator properties in detail, we use the Friend of a Friend (FOAF) ontology. This ontology helps to model properties such as age, gender, education, etc. of the annotator. This is important since this demographic information gives an indication about the reputation of the annotator and helps to form an opinion about particular demographic stereotypes. For example, females provide more descriptive annotations, age groups over 60 provide more annotations, highly educated annotators provide better quality annotations, etc.

1.3.2 Quantifying trust

Once the entities in the annotation process have been modelled, we have to quantify the concept of trust. For this purpose we use probabilistic modelling techniques as well as categorisation techniques.

Probabilistic techniques can be used to build models from positive and negative evidence available and utilise this to predict quality in the future. The evidence is annotations which were provided by annotators during a certain time period in the past and which have been evaluated by professionals at the institutions with a quality score. By defining thresholds on the scores we can categorise the available evidence as positive and negative and use this information to create a reputation profile for the annotators. In our work we use subjective logic to build probabilistic quality and reputation prediction models.

[14] `https://en.wikipedia.org/wiki/Semantic_Web`, August 2015
[15] `http://www.w3.org/TR/prov-o/`

Another method to quantify trust is to utilise the quality categorisation used by the institutions. The institutions have policies and standards regarding quality of annotations provided by annotators based on their relevance to the annotation task and to the artefact. These can be classes such as useful, problematic, judgment, not-useful, etc., with a further sub-division within these classes.

In both techniques we use to quantify trust, we use different properties such as properties of annotator, reputation of annotator and provenance of annotation. We describe in detail in the subsections below about their relevance to trust.

1.3.2.1 Annotator reputation and quality

Quality of annotations is closely linked to the reputation of their annotator. Thus we need to have techniques to model the reputation of annotators. These can be based on evidence on the actual annotation available from the system or on properties of the annotator such as his or her demographics, or on information regarding annotator demographics stereotypes. In our work we model the reputation of annotators based on all three methods.

1.3.2.2 Provenance and quality

The method through which an annotation was produced can give an indication about its quality. This could be information such as timestamp, vocabularies used, etc. We utilise such information to determine trust of annotation and of annotators. For example, annotations produced later during the day might have a lower quality or the length of annotations produced during different intervals of the day or during different days of the week might change. Thus the provenance information becomes relevant to predict quality of annotations. In case of predicting annotator reputation using provenance, the following example demonstrates the importance of provenance. Some annotators participate anonymously and it is difficult to gather sufficient evidence to determine their reputation. We can track an annotator based on patterns in their behaviour such as some annotators login and provide annotations at certain regular intervals of the day or days of the week. These annotations typically have similar patterns in their levels of quality. Also we can cluster patterns to learn how different times have an influence on the quality of annotations.

1.3.3 Predicting trust

After modelling and quantifying trust, the next step is to use this information to predict the quality of annotations and reputation of annotators. We use subjective logic to predict the quality of annotations and also employ machine learning techniques.

1.3.3.1 Subjective logic and semantic similarity for trust predictions

We build annotator profiles using positive and negative evidence from the past and use this to model annotator reputation in subjective logic. We also utilise semantic classification of available annotations into different topics such as annotations about flowers, nature, castles, etc., and utilise this for predicting quality of annotations in the future. The semantic classification is obtained from vocabularies such as Wordnet which are available online. This helps to estimate the reputation of annotators per topic, i.e., their expertise. For annotations belonging to a new topic provided by the same annotator whose expertise topics are already known, we can predict the expertise of this annotator for the new topic based on its semantic closeness to already known topics. The quality of the new annotation can be predicted by weighing the semantic closeness between topics along with the quality of annotations available as evidence.

1.3.3.2 Machine learning techniques for trust predictions

Machine learning techniques help to predict a target value based on various features as input. In our case the target to be predicted is the quality of an annotation or the reputation of an annotator. The features which can be used to predict quality of annotation are its properties such as number of words, parts of speech, presence in vocabularies, etc., along with properties of annotators such as annotator demographics. Also metadata about the artefacts such as creator, title, artefact type etc. are also available from institutions and can be used as features to predict annotation quality. For example, artworks from certain creators are harder to understand and annotate, thereby resulting in lower quality annotations. Thus providing this as a feature helps to make annotation quality predictions.

"Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyse data and recognise patterns, used for classification and regression analysis."[16] We use SVM to perform quality predictions by providing the previously mentioned features as input and obtaining the quality as the predicted target.

[16] `https://en.wikipedia.org/wiki/Support_vector_machine`, August 2015

Since we model the concept of trust using the Semantic Web, we also use prediction algorithms which can exploit the semantic relations between entities for machine learning predictions. Resource Description Framework (RDF) Schema is a language for writing ontologies. There are RDF graph kernels available to perform predictions on RDF graphs. In this case the input is all the features along with the semantic relations between the various features.

The output from the machine learning algorithms is the quality of annotations and the reputation of annotators.

1.4 Objectives and organisation of the thesis

The overall goal of our research is to develop automated or semi-automated techniques to determine quality of crowdsourced annotations for cultural artefacts and reputation of annotators. The main research question is addressed by answering the following research questions:

- How can we model reputation of annotators from the crowd and quality of annotations regarding cultural heritage artefacts?

  In Chapter 3 we propose a workflow which can be employed by cultural heritage institutions to evaluate reputation of annotators and quality of provided annotations. We performed preliminary evaluation of our techniques using subjective logic and semantic similarity measures on the `Steve.Museum` dataset.

- How can different techniques in probabilistic modelling be used to model trust?

  In Chapter 4 we discuss how different operators in subjective logic can be used to model opinions and compare their performance. This is helpful to model opinions when ground truth data is available and we need to make future predictions. In case ground truth data is not available, we discuss about partial evidence. Subjective logic operators can be tuned to model such instances where only partial evidence such as multiple agreements is available.

- How can demographics of annotator and provenance techniques be employed to evaluate the quality of annotations?

  In Chapter 5 we investigate the correlation between different properties of the annotator and the quality of annotation. We also form stereotypes of annotators and

use them for prediction of quality. Apart from annotator demographics we use details about how an annotation was created (provenance) to determine quality of annotations. We later combine the techniques of determining trust based on reputation with techniques employing provenance for determining trust and compare their performance.

- How can efficient techniques be developed for assessing the quality of annotations?

  In Chapter 6 we aim to increase the efficiency of our trust computation algorithms with the goal to maintain or increase the performance while decreasing the computation time. We use machine learning clustering techniques to group semantically similar annotations provided by annotators on the Web about different artefacts and also employ clustering based on provenance information.

- How can machine learning techniques be applied on annotation and annotator features to make predictions on annotator reputation and the quality of annotations?

  In Chapter 7 we determine what is the impact of different annotator demographics such as age, gender, education, etc., and different properties of annotations and provenance of the annotation process on the quality of information. We use machine learning prediction techniques by providing features of annotator, annotation and provenance for training the algorithms. This chapter is built on the initial results we obtained in Chapter 5.

- How can semantic relations and graph properties be combined with machine learning techniques for computing the quality of annotations?

  In Chapter 8 we further extend on our work in Chapters 5 and 7. Instead of using independent properties as features for the machine learning algorithms, we build semantic relation graphs depicting the relation between different entities and then reason on these graphs to determine the quality of annotations. Thus instead of using only the available features regarding the annotator, annotation and annotation process, we also utilise the semantic relationships between these entities and exploit them for machine learning predictions.

1.5 Sources of chapters

The main chapters of this thesis are based on the following publications.

- **Chapter 3**

  - D. Ceolin, A. Nottamkandath and W.J. Fokkink, Automated evaluation of annotators for museum collections using subjective logic, in Proc. 6th IFIP WG11.11 Conference on Trust Management - IFIPTM'12, Surat, IFIP Advances in Information and Communication Technology 374, pp. 232-239, Springer (May 2012)

- **Chapter 4**

  - D. Ceolin, A. Nottamkandath and W.J. Fokkink, Bridging gaps between subjective logic and semantic Web, in Uncertainty Reasoning for the Semantic Web III - Revised Selected Papers of URSW 2011-2013, Lecture Notes in Artificial Intelligence 8816, pp. 242-264, Springer (November 2014)
  - D. Ceolin, A. Nottamkandath and W.J. Fokkink, Subjective logic extensions for the semantic web in Proc. 8th Workshop on Uncertainty Reasoning for the Semantic Web - URSW'12, Boston, CEUR Workshop Proceedings 900, pp. 27-38, CEUR-WS.org (November 2012)

- **Chapter 5**

  - D. Ceolin, P. Groth, A. Nottamkandath, W.J. Fokkink and W.R. van Hage, Analyzing user demographics and user behavior for trust assessment, in Uncertainty Reasoning for the Semantic Web III - Revised Selected Papers of URSW 2011-2013, Lecture Notes in Artificial Intelligence 8816, pp. 219-241, Springer (November 2014)
  - D. Ceolin, P. Groth, W.R. van Hage, A. Nottamkandath and W.J. Fokkink, Trust evaluation through user reputation and provenance analysis, in Proc. 8th Workshop on Uncertainty Reasoning for the Semantic Web - URSW'12, Boston, CEUR Workshop Proceedings 900, pp. 15-26, CEUR-WS.org (November 2012)

- **Chapter 6**

  - D. Ceolin, A. Nottamkandath and W.J. Fokkink, Efficient semi-automated assessment of annotation trustworthiness, Journal of Trust Management 1(1) (May 2014) special issue of PST'13
  - D. Ceolin, A. Nottamkandath and W.J. Fokkink, Semi-automated assessment of annotation trustworthiness, in Proc. 11th Conference on Privacy, Security and Trust - PST'13, Tarragona, IEEE (July 2013). This paper received the Best Student paper award ex-aequo.

- **Chapter 7**

  - A. Nottamkandath, J. Oosterman, D. Ceolin and W.J. Fokkink, Automated evaluation of crowdsourced annotations in the cultural heritage domain, in Proc. 10th Workshop on Uncertainty Reasoning for the Semantic Web - URSW'14, Riva del Garda, CEUR Workshop Proceedings 1259, pp. 25-36, CEUR-WS.org (October 2014)
  - A. Nottamkandath, J. Oosterman, D. Ceolin, A. Bozzon and W.J. Fokkink, Automated evaluation of crowdsourced annotations in the cultural heritage domain, Journal on Data Semantics (Under submission)

- **Chapter 8**

  - A. Nottamkandath, J. Oosterman, D. Ceolin, G.K.D. de Vries and W.J. Fokkink, Predicting quality of crowdsourced annotations using graph kernels, in Proc. 9th IFIP WG11.11 Conference on Trust Management - IFIPTM'15, Hamburg, IFIP Advances in Information and Communication Technology 454, pp. 134-148, Springer (May 2015)

The following publications describe the *Accurator* platform which was developed as part of SEALINCMedia project.

- C. Dijkshoorn, M. Leyssen, A. Nottamkandath, J. Oosterman, M. Traub, L. Aroyo, A. Bozzon, W.J. Fokkink, G.-J. Houben, H. Hovelmann, L. Jongma, J. van Ossenbruggen, G. Schreiber and J. Wielemaker, Personalized nichesourcing: Acquisition of qualitative annotations from niche communities, in Proc. 6th Workshop on Personalized Access to Cultural Heritage - PATCH'13, Rome, CEUR Workshop Proceedings, CEUR-WS.org (June 2013)
- J. Oosterman, A. Bozzon, G.-J. Houben, A. Nottamkandath, C. Dijkshoorn, L. Aroyo, M.H.R. Leyssen, M.C. Traub (2014). Crowd vs. experts: nichesourcing for knowledge intensive tasks in cultural heritage, in Proc. 23rd International World Wide Web Conference, WWW '14, pp. 567-568, Seoul, Republic of Korea, ACM (April 2014)
- J. Oosterman, A. Nottamkandath, C. Dijkshoorn, A. Bozzon, G. Houben, L. Aroyo, Crowdsourcing knowledge intensive tasks in cultural heritage, in Proc. of the 2014 ACM conference on Web science (WebSci '14). ACM, New York, NY, USA, 267-268, ACM (June 2014)

Apart from the above listed publications, the following publications have also contributed to this thesis.

- D. Ceolin, A. Nottamkandath, W.J. Fokkink and V. Maccatrozzo, Towards the definition of an ontology for trust in (Web) data, in Proc. 10th Workshop on Uncertainty Reasoning for the Semantic Web - URSW'14, Riva del Garda, CEUR Workshop Proceedings 1259, pp. 73-78, CEUR-WS.org (October 2014)
- D. Ceolin, P. Groth, V. Maccatrozzo, W.R. van Hage, W.J. Fokkink and A. Nottamkandath, Combining user reputation and provenance analysis for trust assessment, ACM Journal of Data and Information Quality (To appear)

Chapter 2

Preliminaries

In this chapter we present the preliminaries to our work and discuss in detail the different techniques we employed along with the datasets we used in the various chapters.

2.1 Subjective logic

In order to model reputation of annotators and quality of annotations we use a probabilistic logic called subjective logic. This logic is explained in detail in [53] and the reason why subjective logic is a suitable choice for representing trust is explained in detail by Ceolin in [15]. The reasons as to why subjective logic is relevant for modelling trust is as follows. Firstly, subjective logic allows us to represent the truth value of propositions in probabilistic terms, and allows to account for uncertainty in estimation of such value. For the modelling, we require some prior evidence so that trust can be predicted based on this evidence in combination with probabilistic techniques. If no evidence is available, we base our predictions on probability alone. Secondly, it allows to keep track of the subject that made an assertion about the truth value of a proposition. On the Web, data is provided by different sources who have different reliability levels. The ability to keep track of the source that exposes a given piece of data or a subjective opinion is crucial to be able to assess the trust in that piece of data or subjective opinion. Subjective logic allows keeping track of such provenance information and allows reasoning based on the reputation of their source. Thirdly, subjective logic provides a wide range of operators for combining proposition arguments. For instance, operators allow to "discount" an opinion based on the reputation of the source that exposes it, or to compute the truth value (expressed as a subjective opinion) of the logical disjunction or conjunction of two opinions held by the same source.

In subjective logic, arguments are represented by means of so-called "opinions" which are tuples composed of the belief owner or "source", say x, the proposition or "target", say y, and the truth value assigned by the source to the proposition. Subjective opinions are represented as:

$$\omega_y^x$$

or alternatively as:

$$\omega(x : y)$$

Subjective logic makes use of a double probabilistic layer. The probability of each proposition can be represented by means of a binomial distribution (or by means of a multinomial distribution if the proposition is multivalued). A binomial and multinomial distribution are explained as follows.

"A binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p."[1] "A multinomial distribution is a generalisation of the binomial distribution. For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories."[2]

The values of proposition, y, are chosen from among the elements of the set Θ ("frame of discernment"). For instance, if y is a binomial proposition, then $\Theta = \{true, false\}$. A subjective opinion describes the belief in the elements of the power set of Θ (2^Θ). In symbols, an opinion is represented as

$$\omega_y^x(b, d, u, a)$$

when $|\Theta| = 2$ (binomial opinion), or as

$$\omega_y^x(\overrightarrow{B}, u, \overrightarrow{A})$$

when $|\Theta| > 2$ (multinomial opinion). In case of a binomial opinion, b represents the belief in y being *true* and d the belief in y being *false*, i.e., the disbelief. The uncertainty u represents a part of probability mass to which we are unable to assign to either *true* or *false*. a represents the prior probability that y has to be true. In case of a multinomial opinion there is no disbelief because there is no specific *false* value, since y can assume multiple trust values. $\overrightarrow{A}$ represents the vector of prior probabilities for each of the

[1] `https://en.wikipedia.org/wiki/Binomial_distribution`, September 2015
[2] `https://en.wikipedia.org/wiki/Multinomial_distribution`, October 2015

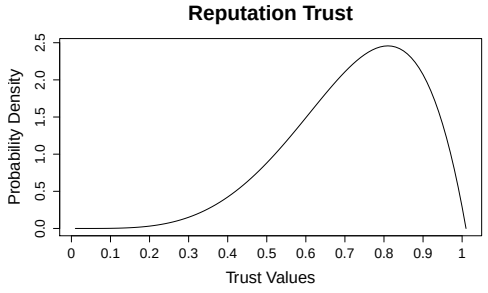

FIGURE 2.1: Example of a Beta probability distribution aggregating four positive pieces and one negative piece of evidence. The most likely trust value is 0.8 (which is the ratio among the evidence). The variance of the distribution represents the uncertainty about the evaluation.

possible truth values of y. $\overrightarrow{B}$ is a vector whose elements are represented as b_x. The values b, d, u are determined by observing pieces of evidence. a is given a priori. In cases where we assume the source to be unknown or implicit, we consider a as $\frac{1}{2}$. The constraints on b, d, u and a are as follows:

$$b \in [0,1] \quad d \in [0,1] \quad u \in [0,1] \quad a \in [0,1] \tag{2.1}$$

$$b + d + u = 1 \tag{2.2}$$

The positive and negative evidence is represented as p and n respectively. The belief (b), disbelief (d), uncertainty (u), and a priori values (a) for binomial opinions are defined by subjective logic as:

$$b = \frac{p}{p+n+2} \qquad d = \frac{n}{p+n+2} \qquad u = \frac{2}{p+n+2} \qquad a = \frac{1}{2} \tag{2.3}$$

The a priori value represents prior probability that the source x knows about proposition y, while belief and disbelief represent the probability mass that x attributes to y being *true* or *false* respectively and uncertainty represents the unassigned probability mass. The value 2 in denominator of a indicates the cardinality of Θ, i.e., the number of values that y can take.

Opinions by a source can be considered in a certain context, i.e. they can be contextualized. For example, source x provides an observation about y in context c (e.g. about an agent's expertise). The most likely value for y in context c, represented as $t(x, y : c)$, is the expected value(E) of the beta distribution corresponding to the opinion and computed as:

$$E = t(x, y : c) = b + a \cdot u. \tag{2.4}$$

A subjective opinion is equivalent to a beta probability distribution for a binomial opinion (as shown in Figure 2.1), which range over the trust levels interval $[0, 1]$ and are shaped by the available evidence, or to a Dirichlet distribution (multinomial opinion). Beta distribution is defined as follows. "Beta distribution is a family of continuous probability distributions defined on the interval [0, 1] parametrised by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution. The beta distribution has been applied to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines."[3]

Dirichlet distribution is the multivariate generalisation of beta distribution. This probability distribution describes the most likely probability values that y can take. If y has Pr probability to be true, since we determine Pr starting from a limited set of evidence, we estimate the most likely value of Pr by means of a beta (or Dirichlet) probability distribution.

One important remark is that this logic allows reasoning on binomial or multinomial data that include for instance URIs. The Beta and the Dirichlet distributions are used because they are "conjugated" [31] with the binomial and multinomial distributions respectively, i.e., their computation is particularly manageable. Other kinds of data and other probability distributions are outside the scope of this logic.

We base our truth estimations on samples of Web data, so the parameter p of the binomial distribution (or the vector of parameters $\overrightarrow{P}$ of the multinomial) is rather uncertain. The Web data sample might be possibly unreliable, uncertain and only partially representative of the entire Web data population. Subjective logic uses a second-order distribution based on the distribution and size of the sample at our disposal to estimate the most likely value that p (or $\overrightarrow{P}$) can take.

Trust is context-dependent, since different users or annotations (or, more in general, agents and artefacts) might receive different trust evaluations, depending on the context from which they situate, and the reviewer. In our scenarios we do not have at our disposal an explicit description of trust policies by the cultural heritage institutions. Also, we do not aim at determining a generic tag (or user) trust level. Our goal is to learn a model that evaluates tags as closely as possible to what the cultural heritage institutions would do, based on a small sample of evaluations provided by them.

[3] `https://en.wikipedia.org/wiki/Beta_distribution`, September 2015

2.1.1 Base-rate discounting operator in subjective logic

An important class of operators of subjective logic is the so-called "discounting" operator. In fact, a subjective opinion allows keeping track of the source of the opinion itself. This permits the reuse of the opinion by third parties, because these third parties, knowing where the opinion comes from, can decide whether to use it. However, before using it, these third parties may require to "smoothen" the opinion to take into account the limited reliability of the source or its possible maliciousness. Therefore, in subjective logic there exists a variety of discounting operators: for instance, one that favors disbelief (to be used if the source is known to be malicious), and one that favors uncertainty (to be used when no specific intention of the source is known). We can also make use of the base-rate sensitive discounting operator in case we just have a probability (i.e., the expected value of an opinion), instead of having at our disposal a complete subjective opinion for a source.

The base-rate sensitive discounting of opinion of source B on y,

$$\omega_y^B = (b_y^B, d_y^B, u_y^B, a_y^B)$$

with the opinion of another source A on B,

$$\omega_B^A = (b_B^A, d_B^A, u_B^A, a_B^A)$$

produces the transitive belief

$$\omega_y^{A:B} = (b_y^{A:B}, d_y^{A:B}, u_y^{A:B}, a_y^{A:B})$$

where the belief b, disbelief d, uncertainty u and apriori a of source A on y based on the opinion of B are computed as follows:

$$\begin{aligned} b_y^{A:B} &= E(\omega_B^A) b_y^B \\ d_y^{A:B} &= E(\omega_B^A) d_y^B \\ u_y^{A:B} &= 1 - E(\omega_B^A)(b_y^B + d_y^B) \\ a_y^{A:B} &= a_y^B. \end{aligned} \tag{2.5}$$

In later parts of this thesis, we use the term "discounting" to refer to the "base-rate discounting" operator in Subjective logic.

2.2 Semantic Web technologies

We convert the datasets available to the Linked Data format. "In computing, Linked Data describes a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried."[4]

Semantic Web technologies include a wide range of formats and technologies aimed at enhancing the Semantic Web vision (which may be summarised with the slogan "moving from a Web of documents to a Web of data"). We use some of them, in particular:

- URIs: Uniform Resource Identifiers offer unique references to any possible entity (e.g.: annotators, artefacts, concepts).
- RDF: the Resource Definition Framework[5] is basically a language for representing graphs. RDF statements are "triples" (Subject, Predicate, Object), where each of these elements can be either a URI or a literal value (with some restrictions).
- Ontologies: defined using RDFS/OWL language, ontologies define types, properties, etc., of URIs in particular contexts. For example, they allow to distinguish URIs referring to sets of users from those representing concepts. We use the following ontologies:
  - Friend Of A Friend[6] (**foaf**): We use this ontology for representing people and connections among them. It models properties of a person such as age, gender, education etc.
  - Simple Knowledge Organization System[7] (**skos**): This ontology is used for representing "concepts" and semantic relations among them.
  - Hoonoh (**hoonoh**) [44]: The Hoonoh ontology provides a vocabulary to represent computed trust metrics. The ontology models person → topic based on three trust factors which are *expertise, experience and impartiality* and models person → person based on trust factors *affinity and track record*. We use this for representing expertise of annotators regarding different topics.

[4] `https://en.wikipedia.org/wiki/Linked_data`, September 2015
[5] `www.w3.org/TR/2002/WD-rdf-concepts-20020829/`
[6] `http://xmlns.com/foaf/spec/`
[7] `http://www.w3.org/TR/skos-primer/`

- RDF Data Cube[8] (**qb**): Statistical datasets comprise of a set of observed values organized along a group of dimensions, together with associated metadata. The Data Cube vocabulary enables such information to be represented using the W3C RDF standard and published following the principles of linked data. Multi-dimensional data can be represented using this ontology. In Chapter 3 we use this ontology to model belief, disbelief and uncertainty parameters in subjective logic.
- Dublin Core Terms[9] (**dcterms**): Various metadata such as creator or title of an artefact can be represented using this ontology.
- W3C PROV Model[10] (**prov**) The PROV Ontology (PROV-O) defines the OWL2 Web Ontology Language[11] encoding of the PROV Data Model[12]. We used this ontology specification to implement provenance applications in cultural heritage use cases. The classes that we used was *prov:Agent* to represent the annotator, *prov:Entity* to represent the annotation and *prov:Activity* to represent the annotation activity. We used properties such as *prov:used*, *prov:wasGeneratedBy*, *prov:startedAtTime*, *prov:endedAtTime* etc., to represent the relation between the different classes.
- Open Annotation (**oa**) [88]: Annotations along with details of artefacts for which it was provided, annotator details and trust values of annotations are modelled using Open Annotation ontology. Currenlty Open Annotation ontology is being re-developed as a W3C standard under the name "Web Annotation Model". However in this thesis we have used the earlier version for modelling purposes.

2.2.1 Converting cultural heritage datasets to Linked Data

The representative dataset comprising annotations or annotator reviews are first transformed into Linked Data, which allows for a uniform representation independent of the original artwork collection and method of acquiring the crowd annotations. Linked Data can also be machine-processed, and external vocabularies and knowledge bases (e.g. DBpedia) in Linked Data format can be easily linked, which helps to obtain more metadata information which can be used as features for machine learning algorithms. In order to transform any dataset into Linked Data, we have to first create structured data using controlled vocabulary terms and dataset definitions represented in the Resource Description

[8] `http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html`
[9] `http://dublincore.org/documents/dcmi-terms`
[10] `http://www.w3.org/TR/prov-o`
[11] `https://www.w3.org/TR/owl2-overview/`
[12] `http://www.w3.org/TR/2012/CR-prov-dm-20121211/`

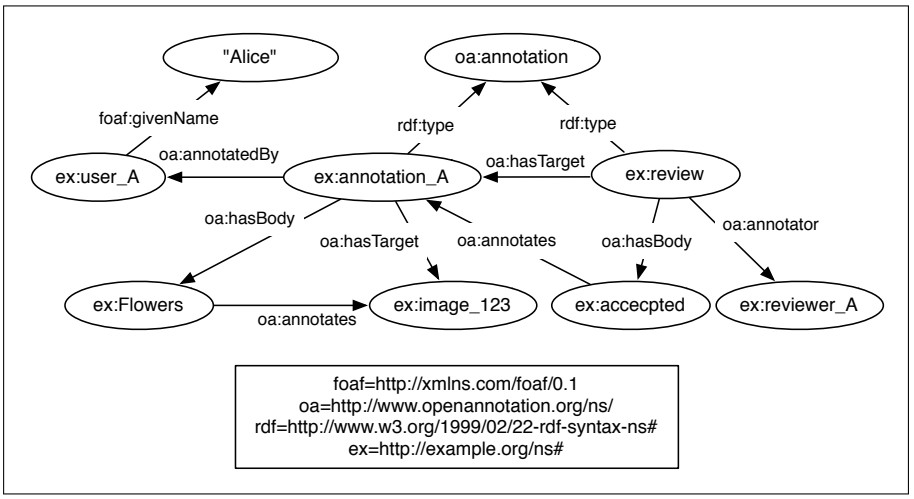

FIGURE 2.2: The figure represents semantic modelling of annotators, annotations and annotation reviews.

Framework serialization formats such as RDFa, RDF/XML, N3, Turtle, or JSON-LD. However, most of the cultural heritage institutions store their data in databases and thus these need to be converted to Linked Data.

Annotations describing artworks provided by the users from the Web are represented using the Open Annotation Model, which helps to link annotations to the user who created them and the artefact for which an annotation was created. A subset of annotations are reviewed by experts at the cultural heritage institutions and their reviews are represented as an annotation of an annotation. A review indicates an expert opinion about the annotation that the user provided according to standards of the institution. Apart from information about annotations, we would like to extend our information about the user who provided the annotation. For users who registered in the system and provided profile information, we model their information using the FOAF ontology, while anonymous users do not have any additional information in their profile. Also the artefact has some meta data such as the creator of the artefact, a title, and material properties. We use the Europeana Data Model (EDM) [45] to represent these properties.

Figure 2.2 shows the semantic model of cultural heritage annotations using standard ontologies.

2.2.2 Enrichment of cultural heritage datasets using external vocabularies

In Chapters 7 and 8 we will enrich the cultural heritage datasets with external vocabularies. The annotations and metadata of the artefacts such as creator name and title are

enriched using external vocabularies such as Wikipedia, DBPedia, ULAN, Wordnet and Flickr. We now describe these vocabularies.

Wikipedia[13] is a mostly unstructured knowledge base maintained by tens of thousands of volunteers worldwide and contains information on a very broad spectrum of topics. The information is intended for human consumption. DBpedia[14] is a semantic repository of information that is extracted from Wikipedia. Most pages on the English Wikipedia have a corresponding entry in DBpedia. Information in DBpedia is structured in RDF and is machine processable. The ULAN is a structured vocabulary maintained by professionals of the Getty Research Institute and contains information such as date of birth and nationality of 647,577 past and current artists (May 2015[15]). For the experiments in this thesis, we had used an older version which consisted of 202,720 artists. *WordNet* is a lexical database for the English language and contains the meaning, synonyms, hyponyms (more general terms), hypernyms (more specific terms), etc. of 147,278 words.[16] Nouns and verbs in *WordNet* are represented using a tree structure and all descend from the top level element *entity*. Adjectives are structured around antonyms (opposite terms) and connected via *similarity* relations.

Flickr is a website where people upload and share their images. Most images are tagged with descriptive labels. Most annotators from the `Steve.Museum` dataset indicated they have tagging experience on Flickr.

The enrichment using these external sources on the `Steve.Museum` dataset that is used in Chapters 7 and 8 is as follows. The `Steve.Museum` dataset contains $1,082$ unstructured creator names. Our goal is to identify creators pointing to individual persons. Therefore we filtered the creator names containing the string *unknown*, locations (countries and places), time periods, and hashed strings to anonymize the details of certain artefacts. This resulted in 742 creator strings (of which some could still point to the same person) which we considered candidate artists. Where possible we put the name in `firstname-lastname` order. We used the preprocessed name to match to DBPedia and ULAN.

Institutions store creator information either as structured, semi-structured or unstructured text. For linking purposes we assume creator text is unstructured. We map ULAN resources using the `gvp:labelPreferred` (e.g. Rembrandt van Rijn) and `gvp:labelNonPreferred` (e.g. Rembrandt Hermanszoon van Rijn) properties. We also map DBPedia resources of type `dbpedia-owl:Artist` using the `foaf:name` property. The textual

[13] `http://en.wikipedia.org/wiki/Wikipedia:About`
[14] `http://wiki.dbpedia.org/About`
[15] `http://www.getty.edu/research/tools/vocabularies/ulan/faq.html`
[16] `https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`

annotations are compared to DBPedia resources based on the `rdfs:label` property to check whether the annotation corresponds to existing words.

For each name that could not be matched we performed a Wikipedia search on that name where we automatically retrieved the top five results and checked if the corresponding DBPedia resources are of the type `dbpedia-owl:Artist`. We automatically made the mapping if there was only one Artist in the results and decided manually when there were multiple Artists. In total 579 candidate artists were mapped onto 479 distinct DBPedia resources. For the ULAN mapping we used both the preprocessed name and the spelling variations on DBPedia if there was a match. In total 470 candidate artists were mapped to 422 distinct ULAN resources. After the mapping process, 442 candidates mapped into both ULAN and DBPedia, 138 only mapped to DBPedia and only 27 mapped to ULAN.

To enrich the annotation we tokenized the annotation and removed stopwords, special characters such as brackets and symbols, and words of length one (single letter annotations). We added a `custom:wikipediaMatchCount` property to each annotation with the number of matched words from the preprocessed annotation. For Flickr we used the `flickr.photos.search` API function searching for the photos containing all annotation words in their label and which were uploaded in 2014. We added a `custom:flickrMatchCount` property to each annotation with the number of photos returned by the API. Finally, to match with the Wikipedia description of the creators we tokenized and stemmed the description, stemmed the preprocessed annotation words and added a `custom:hasCreatorMatchCount` property indicating the number of matched words.

2.2.3 Semantic similarity measure

In order to increase the availability of evidence for our estimate and to let the more relevant evidence have a higher impact on those calculations, we employ semantic relatedness measures as a weighing factor. These measures quantify the likeness between the meaning of two given terms. Whenever we evaluate a tag, we take the evidence at our disposal, and tags that are more semantically similar to the one we focus on are weighed more heavily. Many semantic similarity measures have been developed (see the work of Budanitsky and Hirst [10]). There exist many techniques for measuring semantic relatedness, which can be divided into two groups. First, so-called "topological" semantic similarity measures are deterministic measures of the distance between words based on a word graph (e.g. *WordNet* [72]). Second, there is the family of statistical semantic similarity measures, which includes, for instance, the Normalized Google Distance [21] that measures statistically the similarity between two words on the basis of the number

of times that these occur and co-occur in documents indexed by Google. These measures are characterized by the fact that the similarity of two words is estimated on a statistical basis from their occurrence and co-occurrence in large sets of documents.

We focus on deterministic semantic relatedness measures based on *WordNet* or its Dutch counterpart Cornetto [101]. We use *WordNet* because it is a knowledge graph developed for English language (since `Steve.Museum` which is used mostly for our experiments is in English) and it is freely and publicly available. Also there is an API available to compute semantic similarity measures for the experiments. In particular, we use the Wu & Palmer [107] and the Lin [65] measure for computing semantic relatedness between tags, because both provide us with values in the range $[0, 1]$, but other measures are possible as well.

Among all deterministic semantic similarity measures, our attention focuses on those computed from *WordNet*. It groups words into sets of synonyms called synsets that describe semantic relationships between them. *WordNet* is an acyclic graph where nodes are represented by synsets and edges represent hypernym/hyponym relations. It is a directed graph in which each vertex v is an integer that represents a synset, and each directed edge from v to w represents that w is a hypernym of v. If a synset is a generalization of another one, we can measure the depth, that is the distance between the two. The first ancestor shared by two nodes is the Least Common Subsumer. Since a word can have multiple semantic meanings, we compute the similarity of all synsets of combinations(every combination has one semantic meaning of the word) and pick the maximum value, as we adopt the upper bound of the similarity between the two words. In particular, we use the Wu & Palmer similarity measure [107], which calculates semantic relatedness in a deterministic way by considering the depths between two synsets in the *WordNet* taxonomies, along with the depth of the Least Common Subsumer (lcs) as follows:

$$score(s1, s2) = \frac{2 \cdot depth(lcs(s1, s2))}{depth(s1) + depth(s2)}. \tag{2.6}$$

This means that $score \in [0, 1]$. For deriving the opinions about a concept where no evidence is available, we incorporate *score*, which represents the semantic similarity ($sim(c, c')$) in our trust assessment, where c and c' are concepts belonging to synset $s1$ and $s2$ respectively which represent two contexts. Computation of this score for our experiments is done using an online service[17].

[17] `http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi`

The Lin measure considers the information content of the Least Common Subsumer and the two compared synsets, as follows:

$$2 * \frac{IC(lcs(s1, s2))}{IC(s1) + IC(s2)}$$

where $s1$ is a synset of the first word and $s2$ of the second. IC is the information context, that is the probability of finding the concept in a given corpus, and is defined as:

$$IC(s) = -log\left(\frac{freq(s)}{freq(root)}\right)$$

and *freq* is the frequency of the synset.

So the Wu and Palmer measure derives the similarity of two concepts from their distance from a common ancestor, while the Lin similarity derives it from the information content of the two concepts and their lowest ancestor.

For our earlier experiments, we had used the `Steve.Museum` dataset which had annotations in English. Thus we had used Wu & Palmer semantic similarity measures (which is based on Wordnet where the words are in English) for our trust assessment. In Chapter 6, we worked on SEALINCMedia project experiment dataset which comprised on annotation in Dutch language. For computing semantic similarity for Dutch annotations, we used an implementation provided by the python NLTK library [67] and used pyCornetto[18], an interface to Cornetto, which is the Dutch version of *WordNet*. pyCornetto does not provide a means to compute the Wu and Palmer similarity measure, but it provides the Lin similarity measure, thus for the SEALINCMedia dataset, we adopted the Lin Measure.

2.2.4 Using semantic similarity measure in subjective logic

Subjective logic is well-suited for the management of uncertainty within the Web. As more and more data is added, it becomes important to have modelling techniques which can deal with the uncertainty, and subjective logic is a good candidate. The basis of this logic is the concept of "subjective opinion". It allows to represent how the estimated truth value of an assertion is bound to the source of the corresponding evidence and also helps to easily maintain lightweight provenance information.

In Subjective logic the truth value of assertions is based on the availability of observations. Thus, more the data that is available to compute truth value, the closer we can get to

[18] `https://github.com/emsrc/pycornetto`

the correct truth value for our assertions. In this way subjective logic can benefit from the growing data on the Web.

In order to evaluate trust, it becomes necessary to set the context. This is because it allows delimiting the validity of an opinion and increasing the precision of the corresponding evaluation. For instance, if we gather evidence about the expertise of a user in a given topic, let us say, flowers, then it is important to delimit the validity of the corresponding opinion to the topic "flowers". However, contexts may also impede the use of evidence about a given subject, if the context differs from the context where the evidence was collected. Therefore, we propose to "bridge" contexts by using semantic similarity measures to import evidence from a context to another, after having weighed them on the similarity of the two contexts.

Since we compute opinions based on contexts, it is possible that evidence required to compute the opinion for a particular context is unavailable. For example, suppose that source x owns observations about a proposition in a certain context (e.g. the expertise of an agent about tulips), but needs to evaluate them in a new context (e.g. the agent's expertise about sunflowers), of which it owns no observations. The semantic similarity measure between two contexts, $sim(c, c')$, can be used for obtaining the opinion about an agent y on an unknown or new context through two different methods. We can weigh the evidence at our disposal, and for every piece of evidence, use only the part that corresponds to the semantic similarity between the two contexts. If we have one observation in the known context c' and the similarity between the two contexts is 0.5, then we can use that observation in the new context c as 0.5 piece of evidence.

We introduce here a running example that explains the computation of trust measures using subjective logic and semantic similarity.

Suppose that a user, Alex has contributed three annotations to collection of the Fictitious National Museum: *Buddhist, Indian and Tulips*. Suppose the annotations *Buddhist and Indian* were evaluated as useful and *Tulip* was evaluated as not-useful by the professionals in the museum. Later, Alex contributes a new annotation *Chinese*. If the museum immediately uses this annotation for classifying the artefact, it might be risky because the annotation might be not-useful. Thus our trust computation algorithms rely on a few evaluations of Alex's previous annotations by the Museum, which becomes our ground truth. Based on these evaluations, the system: (1) computes Alex's reputation; (2) computes a trust value for the new tag; and (3) decides whether or not to accept the new annotation.

2.3 User reputation computation and representation

We define a user reputation as a global value representing the user's ability to tag according to the museum policy. By 'global' we mean that the user reputation is not related to a specific context, because this value should represent an overall trust level about the user production: a highly reputed user is believed to have the ability to produce high-quality tags and to choose tags/artefacts related to his/her domain of expertise. Also, the possible number of topics is so high that defining the reputation to be topic-dependent would bring manageability issues. Expertise will be considered when evaluating a single tag, as we will see in the next paragraph.

We require that a fixed number of user-contributed tags are evaluated by the museum. This fixed number then becomes prior evidence for computation of annotators. This number can vary per dataset. For our experiments we have chosen different values for this number. As the value of fixed number of evaluated annotation increases, there will be lesser annotators who have contributed that many annotations and thus it would not be possible to compute reputation for all annotators. It is better to set this number based on average number of tags provided per annotator for that particular dataset and also based on the average number of annotations per annotator which have been evaluated. Based on those evaluations we compute the user reputation using subjective opinions, as in Equation 2.7.

$$\omega_u^m \left(b_u^m, d_u^m, u_u^m, a_u^m\right) = \omega_u^m \left(\frac{p_u^m}{p_u^m + n_u^m + 2}, \frac{n_u^m}{p_u^m + n_u^m + 2}, \frac{2}{p_u^m + n_u^m + 2}, \frac{1}{2}\right) \tag{2.7}$$

where m and u represent the museum and the user, respectively and p and n the count of positive and negative pieces of evidence respectively for computing the values of belief, dis-belief and uncertainty. So, for instance, p_u^m is the count of positive pieces of evidence that the museum m collected about user u, and n_u^m the negative ones.

The algorithm that we will describe makes use of a single value representing the user reputation. The algorithm makes use of the expected value of opinion which is computed using subjective logic as follows.

$$E = b + a \cdot u \tag{2.8}$$

Substituting b, u and a from 2.7, we obtain the expected value of that opinion, as shown in Equation 2.9.

$$E(\omega_u^m) = \frac{p_u^m}{p_u^m + n_u^m + 2} + \frac{1}{2} \cdot \frac{2}{p_u^m + n_u^m + 2} \tag{2.9}$$

To continue with the running example, Alex had annotations *Indian, Buddhist* evaluated as useful and *Tulips* evaluated as not-useful. His reputation is:

$$\omega_{Alex}^{museum} = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}, \frac{1}{2}\right) \quad E(\omega_{Alex}^{museum}) = 0.6 \tag{2.10}$$

2.4 Annotation trust value computation and representation

Tag trust values are represented by means of subjective opinions, as in Equation 2.11.

$$\omega_t^m \left(\frac{p_t^m}{p_t^m + n_t^m + 2}, \frac{n_t^m}{p_t^m + n_t^m + 2}, \frac{2}{p_t^m + n_t^m + 2}, \frac{1}{2}\right) \tag{2.11}$$

Here, we still use the tags created by the user and the corresponding evaluations to compute the trust value, but despite the computation of the user reputation, evidence is weighed with respect to the similarity to the tag to be evaluated. This means that we do not consider each piece of evidence as equally contributing to the computation of the reputation, i.e. evidence is weighed according to the semantic similarity with respect to the tag that we are evaluating. So p and n are determined as in Equation 2.12, where sim is the Wu & Palmer semantic relatedness measure and t is a tag to be evaluated. Despite Equation 2.7, where each piece of evidence is counted as one, here each piece of evidence counts as a real number between zero and one corresponding to the value of the semantic similarity.

$$\begin{aligned} p_t^m &= \Sigma_{t_i \in train} sim(t, t_i) \ \textit{if evaluation}(t_i) = \textit{true} \\ n_t^m &= \Sigma_{t_i \in train} sim(t, t_i) \ \textit{if evaluation}(t_i) = \textit{false} \end{aligned} \tag{2.12}$$

The new annotation *Chinese* inserted by Alex is evaluated as:

$$p_{Chinese}^m = sim(\text{Chinese}, \text{Indian}) + sim(\text{Chinese}, \text{Buddhist}) = 1.05$$

$$n_{Chinese}^m = sim(\text{Chinese}, \text{tulip}) = 0.1$$

$$\omega_{Chinese}^m \left(\frac{1.05}{1.05 + 0.1 + 2}, \frac{0.1}{1.05 + 0.1 + 2}, \frac{2}{1.05 + 0.1 + 2}, \frac{1}{2}\right)$$

$$E(\omega_{Chinese}^m) = 0.95$$

The proposed algorithm is designed so that any other relatedness measure could be used in place of the chosen ones, without the need of any additional intervention. The choice of the semantic similarity and how the semantic similarity is used both affect the uncertainty of the expected results of the algorithms that we propose. In fact, these algorithms use semantic similarity to weigh the importance of evidence when evaluating tags, i.e. words associated with cultural heritage artefacts. We use a deterministic semantic similarity measure which, although it constitutes a heuristics, is based on a trustworthy data source (*WordNet*) and probabilistic semantic similarity measures (Wikipedia similarity measure) for our experiments.

2.5 Statistical tests

In the first subsection we discuss tests which determine the statistical difference between two distributions. Later we discuss about tests to determine correlation between different entities in a dataset.

2.5.1 Statistical distribution tests

Many statistical tests check for normal distribution of datasets. "The normal distribution is remarkably useful because of the central limit theorem. In its most general form, under some conditions (which include finite variance), it states that averages of random variables independently drawn from independent distributions converge in distribution to the normal, that is, become normally distributed when the number of random variables is sufficiently large."[19] For data which does not have a normal distribution, we use the Wilcoxon signed-rank test to observe how statistically different they are. The Shapiro-Wilk test is first used to determine if the test statistic follows a normal distribution.

We used statistical tests to determine how samples in a population differ. We briefly explain these tests.

Sign test In Chapter 3 we obtain annotation quality scores by applying different techniques. In order to observe the statistical difference between the values we use the sign test. "The sign test is a statistical method to test for consistent differences between pairs of observations, such as the weight of subjects before and after treatment. Given pairs of observations for each subject, the sign test determines if one member of the pair tends to be greater than (or less than) the other member of the pair. The paired observations may be designated x and y. For comparisons

[19] `https://en.wikipedia.org/wiki/Normal_distribution`, September 2015

of paired observations (x, y), the sign test is most useful if comparisons can only be expressed as $x > y$, $x = y$, or $x < y$. If, instead, the observations can be expressed as numeric quantities ($x = 7$, $y = 18$), or as ranks (rank of $x = 1st$, rank of $y = 8th$), then the Student t-test or the Wilcoxon signed-rank test. If X and Y are quantitative variables, the sign test can be used to test the hypothesis that the difference between the median of X and the median of Y is zero, assuming continuous distributions of the two random variables X and Y, in the situation when we can draw paired samples from X and Y."[20]

Wilcoxon signed rank test In Chapter 4 we prove that the results from using the weighing and discounting operators in subjective logic for dogmatic opinions are not statistically different by employing the Wilcoxon signed rank test on results from the `Steve.Museum` dataset. In Chapter 5 we use this test to show on a specific dataset that typing duration is statistically different from the *Waisda?* dataset. Also in that chapter we employ the Wilcoxon signed rank test to show that values from our reputation computing algorithms are better than blind guess values and we show that provenance-based techniques and reputation-based techniques are not statistically different. Furthermore, we statistically prove that combining provenance-based and reputation-based estimation techniques provides better results than each of them alone.

In Chapter 6 we use the Wilcoxon signed rank test to compare how the performance of annotators varies during different time intervals (e.g. across days of week, hours of day, weekdays and weekends) in the `Steve.Museum` dataset.

In Chapter 7 we prove that the distribution of parts of speech of annotations provided by the anonymous and registered users in `Steve.Museum` are not statistically different and that frequencies of annotations marked as `judgement` are not significantly different to each other.

"The Wilcoxon signed-rank test [106] is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed."[21]

Shapiro-Wilk test at 95% confidence We use the Shapiro-Wilk test to test normality of distribution in Chapter 6. "The Shapiro-Wilk test is a test of normality in

[20] `https://en.wikipedia.org/wiki/Sign_test`, September 2015
[21] `https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test`, September 2015

frequentist statistics. Normality tests are used to determine if a data set is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed."[22]

Student t-test In some cases we use the Student t-test as an alternative to Wilcoxin signed-rank test if the test statistic follows a Student's t-distribution. We employ this test in Chapter 4 to prove that the weighing and discounting operators in subjective logic are not statistically different.

"A t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution. In probability and statistics, Student's t-distribution is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. Whereas a normal distribution describes a full population, t-distributions describe samples drawn from a full population; accordingly, the t-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution."[23]

Chi-squared test In Chapter 5 we use the chi-squared test to observe the difference in distribution of hour of the day and day of the week in a representative sample and in the whole dataset. "A chi-squared test [80] is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. A chi-square distribution is distribution of a sum of the squares of k independent standard normal random variables."[24]

2.5.2 Statistical correlation tests

In Chapter 5 and 7 we use statistical correlation tests to determine which properties of annotators and annotations are correlated to quality. We decide the tests based on the type of variables used to determine correlation.

[22] `https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test`, September 2015
[23] `https://en.wikipedia.org/wiki/Student%27s_t-test`, September 2015
[24] `https://en.wikipedia.org/wiki/Chi-squared_test`, September 2015

Pearsons chi-squared test In order to determine which categorical properties (e.g. Education) of annotators and annotation are relevant for prediction of quality using machine learning techniques, we use the Pearsons chi-squared test in Chapter 7 on `Steve.Museum` dataset. "Pearson's chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It is suitable for unpaired data from large samples."[25]

Wilcoxon rank sum test For interval properties such as *# words in annotation* we use the Wilcoxon rank sum test to determine correlation with quality in Chapter 7 on the `Steve.Museum` dataset. "Wilcoxon rank sum test is a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis, especially that a particular population tends to have larger values than the other. It can be applied on unknown distributions contrary to student t-test which has to be applied only on normal distributions, and it is nearly as efficient as the student t-test on normal distributions. The Wilcoxon rank-sum test is not the same as the Wilcoxon signed-rank test, although both are nonparametric and involve summation of ranks."[26]

Point bi-serial In Chapter 5 we determine the correlation of annotator profile information with their reputation. For this purpose, we used the point bi-serial correlation metric for categorical variables such as gender on *Wasida?* dataset. "The point biserial correlation coefficient [36] is a correlation coefficient used when one variable is dichotomous, like gender."[27]

Pearson correlation In Chapter 5 we use Pearson correlation to determine correlation of continuous data (e.g. number of tags provided, age) with annotator reputation on the *Waisda?* dataset. "Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables."[28]

[25] `https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test`, September 2015
[26] `https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test`, September 2015
[27] `https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient`, September 2015
[28] `https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`, September 2015

2.6 Machine learning techniques

The goal of our algorithms is to automatically or semi-automatically predict the quality of annotations and the reputation of annotators. In order to perform this task, we should have ground truth data available. The ground truth data is the data which have been evaluated by professionals according to standard policies at institutions and can serve as a good learning example. In order to test our algorithms we split the ground-truth data into two sets, the training set using which we train our algorithms and the test set, on which we test our algorithms and measure the performance of our algorithms.

The splitting of the ground-truth data into training and test sets can be done using n-fold cross validation technique. In one-fold cross validation, we split one part of the ground-truth data (usually around 70%) to the training set and the remaining (30%) data into the test set. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. In Chapter 5 we use one-fold cross validation while in Chapter 7 we use n-fold cross validation. In the following section we describe the various algorithms used in our work in some detail.

2.6.1 Hierarchical clustering algorithm

In Chapter 6 we use hierarchical clustering techniques to group together semantically similar annotations [39]. It is a clustering technique which seeks to build a hierarchy of clusters. In our case the clusters are build based on semantic similarity measures. It is an expensive technique and thus was not used frequently in the experiments. More details about why we use this technique and how it is done for our experiments is explained in Chapter 6.

2.6.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) [22] are very efficient and robust classification algorithms that try to separate classes by finding a maximally separating hyperplane. It is described in detail in [89, 91].

In Chapter 5 we use a regression algorithm to predict the trustworthiness of the annotations. "In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one

understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed."[29] In Chapter 5 we are not interested in predicting the "right" trust value, but rather the class of trustworthiness. Thus we adopt the "regression-by-discretization" approach [62], that allows us to use the SVM to classify our data after having discretized the continuous ones

In Chapter 5, we used the SVM version implemented in the e1071 R library[30].

2.6.3 Naive Bayes

Naive Bayes classifiers, used in Chapter 7, are a family of simple probabilistic classifiers and have been used for text classification in many instances [19, 70]. We identify relevant properties of annotations and annotators and perform predictions using Naive Bayes based on those properties. Naive Bayes classifiers [52] are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. We opted for Naive Bayes classifiers since it has been used extensively for text and document classification and is a simple, yet effective technique. In Chapter 7 we use the Naive Bayes implementation in WEKA[31].

2.6.4 Machine learning using RDF kernels

In Chapter 8 we use machine learning techniques for RDF kernels. Graph kernels have recently evolved into a rapidly developing branch of learning on structured data. They respect and exploit graph topology, but restrict themselves to comparing substructures of graphs that are computable in polynomial time. Graph kernels bridge the gap between graph-structured data and a large spectrum of machine learning algorithms called kernel methods, that include algorithms such as SVMs which were introduced earlier [93].

In Chapter 8 we use the Weisfeiler-Lehman [93] graph kernel for RDF (WLRDF), introduced in [26]. This is a state-of-the-art graph kernel for learning from RDF data in terms of prediction accuracy, with very good computational performance. It uses concepts from the Weisfeiler-Lehman test of isomorphism [105] between subtrees. For each instance, the WLRDF kernel efficiently computes subtree patterns as features, in a number of iterations, where each iteration computes more complex patterns.

[29] `https://en.wikipedia.org/wiki/Regression_analysis`, September 2015
[30] `http://cran.r-project.org/web/packages/e1071/`
[31] `http://cs.waikato.ac.nz/ml/weka/`

Experiments in Chapter 8 have been implemented in Java using the 'mustard' library package of the library at `https://github.com/Data2Semantics/mustard`, which implements different graph kernels, such as the WL RDF kernel for RDF data and wraps the Java versions of the LibSVM [18] and LibLINEAR[32] [30] SVM libraries.

2.7 Evaluation metrics

The measures which we use to determine the performance of our algorithms are briefly explained in this section.

Confusion matrix In the field of machine learning, a confusion matrix is a specific table layout that allows visualisation of the performance of an algorithm, typically a supervised learning one. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa) as shown in Table 2.7. Based on this matrix various scores such as precision, recall, f-measure and accuracy can be computed which give an indication of performance of the algorithms or techniques. We use the terms *TP*, *FP*, *TN* and *FN* for True Positives, False Positives, True Negatives and False Negatives respectively.

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Precision and recall Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

$$Precision = \frac{TP}{TP + FP} \tag{2.13}$$

[32] `http://liblinear.bwaldvogel.de/`

$$Recall = \frac{TP}{TP + FN} \tag{2.14}$$

F-measure In statistical analysis of binary classification, the F-measure is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. The F-measure can be interpreted as a weighted average of the precision and recall, where it reaches its best value at 1 and worst at 0.

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.15}$$

Accuracy Accuracy is used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. It is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{2.16}$$

2.8 Data availability

We use three datasets from the cultural heritage domain in our experiments. The `Steve.Museum` and *Waisda?* dataset are described here. The `SEALINCMedia` dataset is described in Chapter 6.

2.8.1 `Steve.Museum` dataset

`Steve.Museum` [97] is a group of art museums (and professionals who support them) formed in 2006 to explore the role of user-contributed descriptions can play in improving on-line access to works of art. Participants included: Denver Art Museum, Guggenheim Museum, The Cleveland Museum of Art, Indianapolis Museum of Art, Los Angeles Country Museum of Art, The Metropolitan Museum of Art, Minneapolis Institute of Arts, The Rubin Museum of Art, San Francisco Museum of Modern Art, Archives and Museum Informatics, and Think Design. The group was funded in part by the U.S Institute of Museum and Library Services through a National Leadership Grant that ran from October 2006 through September 2008. Due to the different types of art (paintings, sculptures, instruments, etc.) and the fact that the reviewers originate from different heritage institutions (since `Steve.Museum` is a consortium of museums), we believe the `Steve.Museum` is representative for a broad range of cultural heritage collections.

`Steve.Museum` has assembled a test set of works of 1784 works of art, with contributions from all participating museums, and a number of interested, but less active museums. To

enable broad tagging of these works of art, a tagging tool or tagger was made available on the Web, and the tags were recorded in a structured way. The tagger tracks detailed data about registered and anonymous users and the tags they assign , linking tags both to works and to the system environment in which they were given.

Taggers were recruited from the broad Internet community, and asked to tag works of art. Within `Steve.Museum`, taggers were recruited through general museum electronic mailing list requests (e.g. MUSEUM-L), subject-specific lists (such as H-ArtHistory and CAAH), the popular press (including coverage in *The New York Times*) and local press in cities like Indianapolis, blog postings, and volunteer requests on craigslist.org.

Between March 2007 and March 2008, 931 users had registered and there had been an additional 3,949 sessions by unregistered/anonymous users.

The museum professionals also reviewed some of the contributed annotations to determine their quality by using a method similar to that of Von Ahn and Dabbish [100]. They were asked the following question:

*If you found **this work** using **this term** in a **query**, would you be surprised?*

If you are not surprised, then the term can be considered useful. If you are surprised, then the term would be flagged as not useful. Reviews were done at each of the museums, conducted by one or more people. The circumstance of the review was documented in a questionnaire. Each museum group would review all annotations assigned to works from their collection according to common guidelines.

Museum staff indicated, based on the above question, whether the annotations were *Useful* or *Not-useful*. In addition, based on discussions of the `Steve.Museum` team, there were more detailed categorisation. The annotations could be identified as *judgmental*, representing a personal assessment of the art work in a positive or negative way, e.g. "fantastic" or "ugly"; as the result of a *mis-perception*, e.g. a mis-identification of iconography; as a *misspelling* or typo, e.g. "gilrs"; as a reflection of a *personal point* of view or category that the museum can't judge, e.g. "mg2x"; and as a *foreign language term*, e.g. "vert". Based on this, it was possible to determine not only the utility of annotations assigned to works of art, but to qualify the places where annotations might be seen to be not useful. Thus the final list of reviews were as follows: *todo, Judgement-negative, Judgement-positive, Problematic-foreign, Problematic-huh, Problematic-misperception, Problematic-misspelling, Problematic-no_ consensus, Problematic-personal, Usefulness-not_ useful and Usefulness-useful.* Since an annotation could be reviewed by more than one person, there was variability in the reviews.

The majority of the annotations (90%) were evaluated as *useful*. Compared to other crowdsourcing initiatives this was a remarkably good crowd. In experiments where we perform binomial predictions (whether an annotation is of trustworthy or not), we consider only *Usefulness-useful* as a positive evaluation, all the others are considered as negative evaluations. The tags classified as *todo* are discarded, since their evaluation has not been performed yet.

The `Steve.Museum` dataset is provided as a MySQL database and consists of several tables. Those most important for us are: "steve_term" that contains information like the identifiers for the artefact annotated and the words associated with them (annotations); "steve_session" that reports information about when the annotations were provided and by whom, and "steve_term_review" that contains information about the annotation evaluations. We join these tables and we select the information that is relevant for us: the annotations, their annotators, their timestamps (i.e. date and time of creation) and their evaluation.

For our predictions we partition the `Steve.Museum` dataset published online which contains the evaluated annotations into a training and a test set. We use the training set to learn a model for evaluating annotations. We perform predictions of evaluations of annotations in the test set based on our trust model.

The `Steve.Museum` project also allowed users to register and enter personal information, which 488 users (40%) did. The characteristics that could be entered are shown in Table 2.2. The *community* property relates to the area of living of the user and differentiates between "rural", "suburban" and "urban". *Experience* is the experience the user has with art in general with options "novice", "intermediate" or "expert". *Education* is the highest level of education of the user, either "secondary or high school", "some college or college graduate", "advanced degree" or "advanced degree in art history". *Household income* had four options ranging from "less than 30k/year" to "75k/year or more". Users could indicate whether or not they *worked in a museum* and how many *museum visits* they had over the years. Users could indicate their *involvement level* with artwork annotation by selecting "not active", "somewhat active" or "very active". *Tagging experience* indicates whether or not the user had any previous experience with tagging. *Internet connection* provides four options ranging from "dial-up" to "cable" and *internet usage* provided six options ranging from "every few weeks" to "very active".

In Chapter 7, in order to train the annotation quality classifier, each annotation must be labeled with a single judgement. However, from the total of 44,448 annotations, a detailed analysis performed in this chapter showed that the `Steve.Museum` dataset contains 8500 annotations that were judged by multiple reviewers. We employed a simple reconciliation technique, by keeping only unanimous judgements, therefore excluding

4571 annotations. This relatively high number of conflicts is at least partly due to the fact that there are many fine-grained verdicts an annotator can give (for e.g. judgement-positive, judgement-negative, problematic-foreign etc.) and annotations were reviewed by experts and their opinion had to be weighed equally. In total only 30 annotations were reviewed as *judgements*, too few for training purposes, hence we removed them from the dataset. Table 2.1 contains a summary of the dataset.

TABLE 2.1: Summary of the Steve.Museum dataset.

Description	*Values*	
Number of annotators / Registered	1.218 / 488	(40%)
Total annotations	44.448	
Unique annotations	13.099	
Annotations evaluated as *useful*	40.012	(90%)
Annotations evaluated as *not useful*	4.227	(9%)
Annotations evaluated as *problematic*	209	(1%)

TABLE 2.2: Annotator properties in the $\boldsymbol{F_u}$ feature set and the percentage of registered annotators who filled in the property.

Features	*# of users (%)*	
community	431	(88%)
experience	483	(99%)
education	483	(99%)
age	480	(98%)
gender	447	(92%)
household income	344	(70%)
works in a museum	428	(88%)
museum visits	370	(76%)
involvement level	411	(84%)
tagging experience	425	(87%)
internet connection	406	(83%)
internet usage	432	(89%)

The dataset contains contributions both from users that provided their demographic data (we call them "registered users") and users who did not provide any personal detail (we refer to these as "anonymous users"). Table 2.2 lists the annotator properties and the percentage of registered users who filled in each property. The only information that we have at our disposal for the latter is an anonymised hash value of their email address. Several users share the same email address hash. However, these coinciding hash values might be due to the use of a default email address instead of the users' personal one. Hence we do not assume that the anonymous users that share the same hashed email address coincide.

We characterize and compare the contributions from the two groups of users, in order to identify peculiarities, differences and similarities, if present. First of all, we observe

a difference in the size of the two contributions, as the contributions from anonymous users represent 43% of the total amount of annotations.

The two sets of contributions are similar with respect to the number of words that compose them: 1.29 in the case of annotations from anonymous users, 1.30 for the registered ones. Likewise, the average length of the words that compose the annotations are very similar: average word length for the annotations provided by anonymous users is 6.24 characters, while by registered users it is 6.39 characters.

We analysed the parts of speech distribution of the annotations from the two groups of users. We utilised the OpenNLP R library[33] to interface to Apache OpenNLP[34] to identify the part of speech that corresponds to each word in the annotations. The distribution of the parts of speech of the annotations provided by the anonymous and registered users is not significantly different, as proven by a Wilcoxon signed-rank test at 95% level of significance. These distributions indicate for instance that 71.0% of the contributions provided from anonymous users are nouns, 6.0% are verbs and 15.3% are adjectives, while the contributions from registered users are 70.8% nouns, 6.3% verbs and 15.1% adjectives.

There is a high similarity also concerning the average performance per session between registered and anonymous users as shown in Table 2.3.

TABLE 2.3: Comparison of the average performance per session between registered and anonymous users.

Evaluation Category	**Average frequency per session (Registered users)**	**Average frequency per session (Anonymous users)**
usefulness-useful	75.57%	74.46%
usefulness-not_useful	11.19%	11.96%
problematic-personal	0.53%	0.61%
problematic-no_consensus	0.69%	0.63%
problematic-foreign	0.99%	1.13%
problematic-huh	0.36%	0.55%
problematic-misperception	2.65%	2.76%
problematic-misspelling	0.88%	0.89%
judgement-positive	0.70%	0.48%
judgement-negative	0.75%	0.95%
comments	2.15%	1.72%
not evaluated	3.54%	3.86%

The only significant difference that we identified between the annotations provided by the two groups regards the time of the day and the day of the week when these were

[33] `http://cran.r-project.org/web/packages/openNLP/index.html`
[34] `https://opennlp.apache.org/`

contributed. In particular, the contributions from anonymous users present a higher frequency of occurrence in the first hours of the day (00:00 - 04:00), while the contributions from registered users are spread over the rest of the day with higher frequency. Likewise, the contributions from anonymous users are more frequent from Thursday to Tuesday, while those from registered users are more frequent in the central part of the week, from Tuesday to Thursday.

In Chapters 3 and 4 we use algorithms to determine quality of annotations by building reputation models for annotators employing subjective logic and semantic similarity measures. For the experiments in these chapters, we had 44,448 annotations.

In Chapters 7 and 8 we focus on building prediction algorithms at annotation level. Once we performed this grouping we realised that some of the annotations had multiple reviews by museum professionals. We discuss in Chapter 7 how we dealt with such scenarios. Also we perform enrichment of the `Steve.Museum` dataset in these chapters; the statistics are described in the next section. Thus the statistics of the `Steve.Museum` dataset used in Chapters 7 and 8 are as listed in Table 2.4.

2.8.2 Enriched `Steve.Museum` dataset

We convert the `Steve.Museum` dataset into linked dataset as per techniques described in Section 2.2.1 and perform enrichment using techniques listed in Section 2.2.2. Table 2.4 provides a summary of the complete linked and enriched dataset. The transformed dataset and the enrichments have been performed and the output is in RDF/XML files.[35]

TABLE 2.4: Summary of the transformed and enriched `Steve.Museum` dataset.

Total number of triples	473,986	
Annotators / registered annotators	1,218 / 488	(40%)
Annotated artworks	1,784	
Candidate creators / mapped creators	1,082 / 605	(56%)
Annotations in Flickr (>0 images retrieved)	25,591	(56%)
Annotations in DBpedia (>0 words matched)	25,163	(55%)

A more in-depth analysis of the dataset is provided in Chapter 7.

2.8.3 Waisda? dataset

Waisda?[36] is a video tagging gaming platform launched by the Netherlands Institute for Sound and Vision in collaboration with the public Dutch broadcaster KRO. The game's

[35] Dataset available at `https://github.com/anottamkandath/Datasets/tree/master/Chapter7`.
[36] http://www.waisda.nl

logic is simple: users watch video and tag the content. Whenever two or more players insert the same tag about the same video in the same time frame (10 seconds, relative to the video), they are both rewarded. The number of matches for a tag is used as an estimate of its trustworthiness. When a tag is not matched by others, it is not necessarily considered to be untrustworthy, because, for instance, it can refer to an element of the video unnoticed by other users, or it can belong to a niche vocabulary. Thus, tags that have no matches are not necessarily wrong. In the game, when counting matching tags, typos or synonymities are not taken into consideration. This dataset was also provided as a MySQL database and consists of several tables with details of annotator profile, video metadata, provenance information and details of matched tags. However, the information in user profiles in this dataset is much less compared to the `Steve.Museum` dataset and the annotations are provided in Dutch language. Nonetheless, this dataset is useful to demonstrate that the techniques we develop can be applied to a dataset in another language and to understand behaviour of annotators who are more regionally clustered.

Chapter 3

Trust Evaluations using Subjective Logic and Semantic Similarity

In this chapter we present the annotation process of cultural artefacts. Techniques to represent, model and predict quality of annotations are employed which utilise annotator reputation and expertise. The work in this chapter introduces the problem of trust for cultural artefacts and different directions of work in following chapters. This chapter is based on a paper presented at the 6th IFIP Trust Management Conference(IFIPTM 2012) in Surat, India. My contributions are designing the workflow, helping to design the algorithm, performing part of the experiments and part of the manual and empirical evaluations.

3.1 Introduction

Annotations for cultural artefacts should be provided only by trusted sources, and should be validated by museum professionals or users who have proven sufficient expertise in the same topic. However, in the real world, annotations are provided by users from the Web and it is hard to know about their trustworthiness beforehand. Also museum professionals do not have enough time or expertise to manually evaluate the provided annotations. We assume that a generic initial classification of an artefact is already available in the form of a specific set of tags or keywords provided by professionals from the museum (e.g. indicating the period of production or the type of artefact). Most museums use a standard thesaurus (such as Iconclass[1]), which serves as a basis for deriving relations between the various artefacts and forms a controlled vocabulary for annotations. In our evaluations, we show that semantic similarity measures in combination with subjective

[1] `http://www.iconclass.org/`

logic are quite helpful in predicting quality of annotations. Semantic similarity measures can also help in selection of annotators who can provide an annotation for a certain topic based on a proper average of an annotator's reputation and the "semantic similarity" between the requested topic and the recorded expertise areas of the annotator. Thus trustworthy annotators can be selected to provide quality annotations.

Moreover, we employ mechanisms for evaluating the quality of provided annotations. We envision a workflow which can be employed by cultural heritage institutions to constantly manage and update the trust, reputation and expertise information of registered annotators by employing trust algorithms based on subjective logic which is described in Chapter 2. This work extends previous work on determining the quality of annotations using (Semantic) Web sources [17] by combining subjective logic with measures of semantic relatedness, thereby providing an extensive model for managing annotations.

3.2 Workflow of annotation evaluation

Our model aims at obtaining trustworthy annotations through crowdsourcing. We present the workflow at cultural heritage institutions from gathering the annotations to estimating their quality. We then discuss a data representation of subjective opinions and the algorithm to determine trustworthy annotations. These two parts are connected by subjective opinions: the first part provides a representation for the expertise, i.e., the "object" of our opinions, whereas the algorithm computes the trust levels and outputs the most trustworthy annotations.

The annotators can register at the cultural heritage institutions website by providing their personal information along with topics they are interested in to annotate or in which they believe they have sufficient expertise. These topics can be the titles of different collections such as flowers, castles, etc., as categorised and listed by the institutions. The institutions can then verify their expertise level through various techniques such as an initial questionnaire. The institutions can then create annotations tasks for artefacts which are not sufficiently described in their collection. Based on the list of topics provided in the annotator profile, the institutions can select a list of annotators to perform annotation tasks. This selection can also incorporate semantic similarity between topics, so that annotators who are experts on a particular topic can be recommended for semantically similar topics. The annotations provided are evaluated for their quality by professionals at these institutions and the annotator profile is updated based on their contributions and expertise level in different topics. The updated expertise profile is used to recommend artefacts for annotation in the future. The proposed workflow is listed in Figure 3.1.

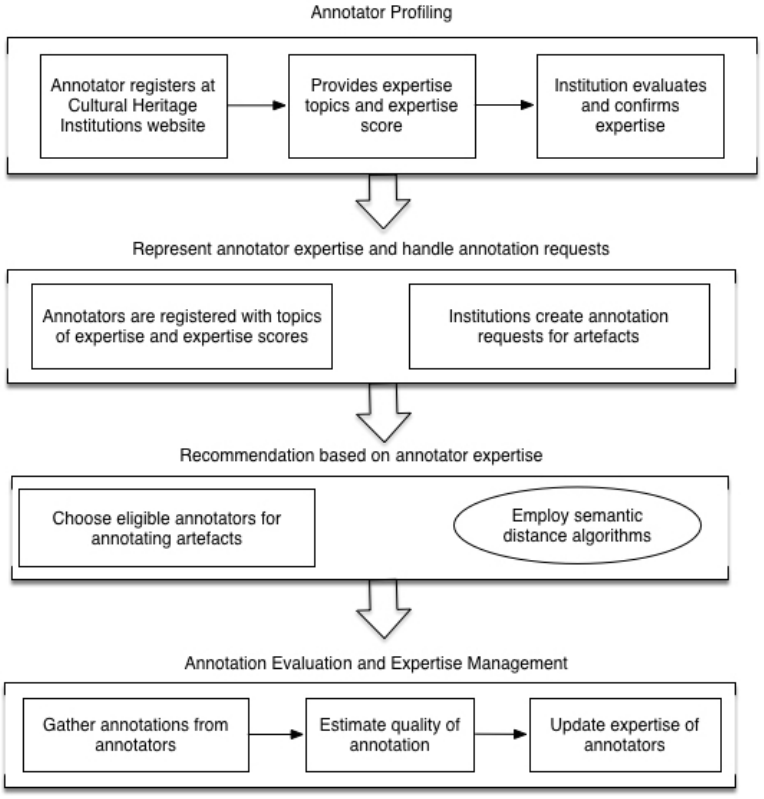

FIGURE 3.1: Workflow.

3.2.1 Data representation

The Open Annotation model presented in Chapter 2 can be used to model the annotations along with annotator details. Here we provide an overview of other modelling techniques to model the trust values of annotations and expertise of annotators.

The Hoonoh ontology provides a vocabulary to represent computed trust metrics. The expertise (for e.g. E1) of each annotator is recorded, through the hoonoh ontology, by linking the URI representing the annotator (modelled as foaf:person) to the URI representing the topic of expertise (for e.g. T1). The *ExpertiseRelationship* is used to represent person $\rightarrow$ topic relationship. We use Hoonoh ontology to represent the relationship between an annotator and topic.

The reason why we use this ontology is because it maps to existing vocabularies such as FOAF and SKOS, and when combined with appropriate algorithms it can be used to populate the Semantic Web with data to support expert finding initiatives. These initiatives can be quite useful to cultural heritage institutions to find the best annotators from the Web for performing annotation tasks.

We show examples of how this can be represented using RDF statements by the institutions:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ex: <http://example.org/ns#>
PREFIX hoonoh: <http://hoonoh.com/ontology#>
ex:T1 a hoonoh:Topic, skos:Concept.
ex:user rdf:type foaf:Person .
ex:E1 a hoonoh:ExpertiseRelationship;
    hoonoh:from ex:user;
    hoonoh:toTopic ex:T1 .
```

Listing 3.1: Representing the annotator expertise by means of the hoonoh ontology

Data Cube Vocabulary allows to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts. We define a data structure using RDF datacube representing a subjective opinion, link it to the corresponding hoonoh:ExpertiseRelationship, and then populate it with observations, i.e., opinion instances:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ex: <http://example.org/ns#>
PREFIX hoonoh: <http://hoonoh.com/ontology#>
PREFIX prov:  <http://www.w3.org/ns/prov#>
PREFIX qb:  <http://purl.org/linked-data/cube#>
PREFIX dcterms:  <http://purl.org/dc/terms/>

ex:Opinion rdf:type qb:DataStructureDefinition;
    qb:component
        [ qb:measure ex:belief; ],
        [ qb:measure ex:disbelief; ],
        [ qb:measure ex:uncertainty; ],
        [ qb:measure ex:apriori; ] .

ex:dataset rdf:type qb:DataSet;
    qb:structure ex:Opinion;
    dcterms:subject ex:E1 .

ex:obs1a rdf:type qb:Observation, prov:Entity;
    qb:dataSet ex:dataset;
    prov:wasAttributedTo ex:Museum;
    ex:belief 0.4;
    ex:disbelief 0.2;
    ex:uncertainty 0.4;
    ex:apriori 0.5.
```

Listing 3.2: Representing a subjective opinion about the annotator expertise

3.2.2 Trust (expertise) management

We are interested in determining the annotator expertise about a given topic, so, if **ex**:E1 is of type **hoonoh**:ExpertiseRelationship, an opinion is:

$$expertise(user, T1) = \omega^{\textbf{ex}:\text{E1 } \textbf{hoonoh}:\text{from } \textbf{ex}:\text{user}}_{\textbf{ex}:\text{E1 } \textbf{hoonoh}:\text{toTopic } \textbf{ex}:\text{T1}}(b, d, u, a) \tag{3.1}$$

The institutions can determine the initial level of annotation expertise for an annotator through different techniques. For e.g. a questionnaire covering annotation tasks for different artefacts in their collection where ground truth is already available can be set before the actual annotation tasks. An annotators performance in this questionnaire can be used as evidence for building their reputation and in subjective logic it can become the *apriori* component, which provides an initial indication of the annotator expertise. If no such technique can be applied to get an initial idea of the annotators expertise, the *apriori* component can be set to 0.5. As the user provides candidate values for annotations and these are evaluated, the weight of the *apriori* on the trust value will decrease.

When evaluating the expertise of the user about a topic T1, the opinion is computed using subjective logic as mentioned in Chapter 2, The aggregated opinion is computed by weighing semantic similarity of each piece of evidence with regard to T1.

3.2.3 Algorithm

Algorithm 1: Algorithm to predict trustworthy annotations

```
1 for request ← request_1 to request_n do
2 |  users ← select_users(request)
3 for user ← users_1 to users_n do
4 |  result ← append_value(user, request)
5 output ← evaluate_results(result)
6 update_expertise(users)
7 return output;
```

We introduce in Algorithm 1 a pseudo-code description that computes trust levels and outputs the most trustworthy annotations, and we provide a qualitative description of it. We explain the functions employed in Algorithm 1.

select_users This function selects a set of annotators to whom we forward an input *request*. A *request* should contain:

- A reference to the artefact to be annotated.

- A first, high-level classification of the item, that facilitates the annotators selection (e.g., the century when it was made)
- The requested "facet", necessary to obtain comparable candidate values (e.g., the "what" facet, i.e. the artefact content).

The selection procedure depends on the cultural heritage institutions, so we do not make it explicit. Some examples can be as follows:

- Select n annotators with the highest ranked expertise about a requested topic (which can be decided either through the questionnaire or based on their performance for annotation tasks)
- Consider all the experts. Weigh their reputation with regards to the distance from the request. Order and select them.
- Consider also the belief and uncertainty (and impose some conditions on them) when selecting annotators.

append_value Collects the contributions obtained from the selected annotators. *result* is a list of couples like (*value*, *annotators_opinions*) where *value* is the annotation and *annotators_opinions* is the known opinion about that annotator.

evaluate_results Aggregates results and takes a decision about them. There are many operators in Subjective logic to perform this operation. For example, we could use the cumulative fusion operator [56] or the consensus operator [54] as a possible aggregation function. A possible decision strategy is to choose the highest-rated value. A decision strategy has to select a candidate value, while reducing the risk of taking a wrong decision and solving possible controversies, such as when multiple candidate values all share the highest rank.

update_expertise After having evaluated the candidate values for the annotation, annotators will be "rewarded" (if their candidate was selected) or "penalized" (otherwise). In principle, this means incrementing positive evidence if the previously evalutated annotation was considered useful, else incrementing negative evidence if the annotation was not considered useful.

output The annotation selected can be directly accepted by the museum, or ranked qualitatively according to its trust level (e.g. "accept" when trust level is higher than 0.9, "review" otherwise), so that appropriate actions are taken.

3.3 Annotation Evaluation

This section describes some analyses performed on the `Steve.Museum` dataset, which is described in detail in Chapter 2, for validating our proposed approach. The goal of our evaluations is twofold. First we want to show that semantic similarity measures are indeed useful for predicting quality of annotations and that annotators who provide annotations belonging to a particular broader topic, also tend to provide other annotations with similar quality and belonging to the same broader topic. Secondly we build a trust model using subjective logic and semantic similarity measures and try to show that we are able to predict the quality of annotations.

In order to achieve our first goal, we manually go through the `Steve.Museum` dataset and observe semantically similar annotations which are provided a lot of times by different annotators. For this experiment, we computed the semantic relatedness by using the Wu & Palmer measure on *WordNet*. One such example of a semantic cluster is shown in Figure 3.2, where the annotations *Chinese, Asian and Buddhist* are provided by the annotators shown in the graph. From the graph we can see that an e.g. annotator with ID 2380 provided around 23 annotations of *Chinese* which were evaluated as useful, and around 17 annotations of *Asian* and 7 annotations of *Buddhist*, both of which were also evaluated useful by professionals). In the graph, an annotator with ID 2382 provided not-useful annotations of all three *Chinese, Buddhist and Asian* along with useful ones.

Another cluster that we considered had annotations *Piano, Music, String and Instrument* and we had many other such clusters. We compared the expertise of annotators using annotations from those clusters and noticed that people having a high amount of positive (or negative) evidence regarding an annotation in a particular cluster also had a high amount of positive (or negative) evidence about the other annotations in the same cluster. Positive and negative evidence is derived from the evaluation by the museum: annotations evaluated as useful are counted as positive evidence, while non-useful ones are considered as negative evidence. This manual and empirical analysis gave us a first concrete evidence about the relatedness between reputation based on evidence and semantic similarity.

We also computed reputation of annotators using subjective logic techniques as mentioned in Section 2.3 of Chapter 2 and computed quality of annotations[2] by using the technique of weighing semantic similarity measures and evidence as described in Section 2.4 of Chapter 2. The predictions of the annotation quality is a value in interval [0,1]. We set a threshold to label annotations with trust values of at least 0.7 as "useful", while others as "not-useful". Weighing the semantic similarity measure on the evidence

[2]Analyses is available at: `https://github.com/anottamkandath/Datasets/tree/master/Chapter3`

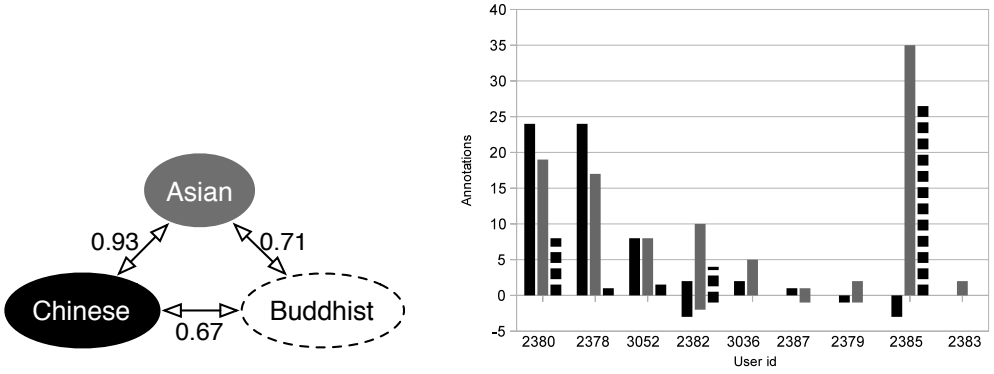

FIGURE 3.2: Cluster and corresponding positive/negative evidence per user.

improved the performance of subjective logic in a statistically significant manner, as proven by applying the sign test with a confidence interval of 95% on the compared errors. We also would like to point out that as a consequence of weighing, *uncertainty* of reputations rises, since weighing reduces the amount of evidence considered. However, often this consequence did not worsen our results, especially when the reputation of annotators were already quite high (e.g., the reputation of an annotator reduced to 0.92 from 0.97).

3.4 Conclusion

We demonstrated the potentials of combining subjective logic, semantic relatedness measures and Semantic Web technologies for handling users expertise and annotations trustworthiness. We presented a workflow, data representation technique and algorithm to evaluate quality of annotations for cultural artefacts.

Chapter 4

Using Subjective Logic for Computing Trust

In this chapter we discuss how to extend and apply subjective logic techniques to compute trust opinions. The work in this chapter was first presented at the 8th International Workshop on Uncertainty Reasoning for the Semantic Web at the 11th International Semantic Web Conference (ISWC 2012) in Boston, USA, and later published in Uncertainty Reasoning for the Semantic Web III - Revised Selected Papers of URSW 2011-2013, Lecture Notes in Artificial Intelligence 8816, Springer LNAI 8816 Proceedings in 2014. For the section Combining subjective logic with deterministic semantic similarity measures, I investigated how the different operators in subjective logic can be combined with deterministic semantic similarity measures and performed the experimentation and evaluation. For the section Combining probabilistic semantic similarity measures within subjective logic, I investigated different probabilistic semantic similarity measures and proposed to use the Wikipedia semantic similarity measure.

4.1 Introduction

In Chapter 2 we described subjective logic and how it can be combined with semantic similarity measures for weighing opinions. In this chapter we propose extensions and applications of subjective logic with semantic similarity, namely: the use of deterministic and probabilistic semantic similarity measures for weighing subjective opinions and a method for accounting for partial observations. The deterministic semantic similarity measure is computed using the Wu & Palmer measure from *Wordnet* graph and

is evaluated on the `Steve.Museum` dataset, while the probabilistic one is computed using Wikipedia semantic similarity measure and has been provided with a mathematical underpinning.

In some cases there is no ground truth available from cultural heritage institutions to build our algorithms for trust. However, multiple annotations from different annotators may have been collected as evidence for the same artefact, and they can be considered as partial evidence. In this chapter we introduce a new technique for the validation of partial observations. It has been employed on examples from the *Waisda?* dataset.

4.2 Related work

The core element of subjective logic is the concept of "opinion", that is, the representation that a given source holds with respect to the truth value of a given proposition. Subjective logic's operators allow combining opinions in different manners, and their development has been widely investigated. Remarkably, the averaging and cumulative fusion operators [55, 56] (i.e., which allow averaging or cumulating opinions about the same proposition from different sources) and the discounting operator [57] (i.e., the operator that allows weighing a source's opinion based on the source's reputation) operators are among the most generic and useful operators for this logic. The weighing and discounting based on semantic similarity measures can resemble the work of Jøsang et al. [55], although the additional information that we include in our reasoning (that is, semantic similarity) is related only to the frame of discernment in subjective logic, and not to the belief assignment function. These operators provide the foundations for the work proposed in this chapter. The exploration of uncertain partial observations used for building subjective opinions has been done by Kaplan et al. [59]. Unlike their work, we restrict our focus on partial observations of Web-like data and evaluations, which comprise the number of "likes", links and other similar indicators related to a given Web item.

4.3 Combining subjective logic with deterministic semantic similarity measures

The semantic similarity measures are basically split into two main classes: deterministic and probabilistic semantic similarity measures. The deterministic ones are based on deterministic computations made, for example, on word graphs (e.g. *WordNet*) as

mentioned in Chapter 2. The probabilistic ones apply probabilistic reasoning to derive semantic relatedness between words based, for instance, on the occurrence and co-occurrence of these two words in large document corpora. We extend subjective logic to incorporate these measures by representing semantic similarity measures by means of subjective opinions (discounting), or by using the similarity measures to weigh items of evidence before using them to build subjective opinions (weighing).

4.3.1 Using deterministic semantic similarity measures within subjective logic

The method of weighing opinions is described in detail in Chapter 2. Here we describe in detail the discounting of opinions in subjective logic.

We can compute a subjective opinion in the known context c' and then use it in the unknown context c after having "discounted" it (using the subjective logic discounting operator). The discounting factor would be a subjective opinion that represents the semantic similarity between the two contexts. The reason why we have these two different approaches is that weighing operates directly on the evidence, while discounting applies on the subjective opinion. In the latter case uncertainty has already been quantified (and therefore some probability mass has been assigned to it), while in the first case not. Hence, the choice between the two alternative depends on the strategy chosen (it could be that operating on the opinion is more computationally efficient, and hence discounting is preferable), or on case study constraints (e.g., if the evidence from a given context are already expressed as a subjective opinion, then it is simpler to use it than to revert it to the pieces of evidence on which it is computed). In opinion discounting, every opinion source x has about other related contexts c', where $c' \in C$, is discounted with the corresponding semantic similarity measure $sim(c, c')$ using a discounting operator in subjective logic. The discounted opinions are then aggregated to form the final opinion of x about y in the new context c.

Subjective logic offers a variety of operators for "discounting", i.e. for smoothing opinions given by third parties, provided that we have at our disposal an opinion about the source itself. "Smoothing" is meant as reducing the belief provided by the third party, depending on the opinion on the source (the worse the opinion, the higher the reduction). Moreover, since the components of the opinion always sum to one, reducing the belief implies an increase of (one of) the other components: hence there exists a discounting operator favouring uncertainty and one favouring disbelief. Finally, there exists a discounting operator that makes use of the expected value E of the opinion. Following this line of thought, we can use the semantic similarity as a discount factor for opinions imported

from contexts related to the one of interest, in case of a lack of opinions in it, to handle possible variations in the validity of the statements due to a change of context.

So, we need to choose the appropriate discounting operator that allows us to use the semantic similarity value as a discounting factor for opinions. The disbelief favouring discounting is an operator that is employed whenever one believes that the source considered might be malicious. This is not our case, since discounting is used to import opinions owned by ourselves but computed in different contexts than the one of interest. Hence we do not make use of the disbelief favouring operator.

In principle, we would have no specific reason to choose one between the uncertainty favouring discounting and the base rate discounting. Basically, since only rarely the belief (and hence the expected value) is equal to 1, the two discounting operators decrease the belief of the provided opinion, one by multiplying it by the belief in the source, the other one by the expected value of the opinion about the source. In practice, we will see that, thanks to Theorem 4.1, these two operators are almost equivalent in this context.

Theorem 4.1 (Semantic relatedness measure is a dogmatic opinion)**.** *Let* $sim(c, c')$ *be the semantic similarity between two contexts* c *and* c' *obtained by computing the semantic relatedness between the contexts in a graph through deterministic measurements (e.g. [107]). Then,* $\forall\ sim(c, c') \in [0,1]$*,*

$$\omega_{c=c'}^{measure} = (b_{c=c'}^{measure}, d_{c=c'}^{measure}, u_{c=c'}^{measure}, a_{c=c'}^{measure})$$

is equivalent to a dogmatic opinion in subjective logic, i.e., a subjective opinion with uncertainty equal to zero.

Proof. A binomial opinion is a dogmatic opinion if the value of uncertainty is 0. The semantic similarity measure can be represented as an opinion about the similarity of two contexts c and c'. However, since we restrict our focus on *WordNet*-based measures, the similarity is inferred by graph measurements, and not by probabilistic means. This means that, according to the source, this is a "dogmatic" opinion, since it does not provide any indication of uncertainty: $u_{c=c'}^{measure} = 0$. The opinion is not based on evidence observations, but rather on actual deterministic measurements.

$$E(\omega_{c=c'}^{measure}) = b_{c=c'}^{measure} + u_{c=c'}^{measure} \cdot a = sim(c, c'), \tag{4.1}$$

where *measure* indicates the procedure used to obtain the semantic relatedness score, e.g. the Wu and Palmer measure. The values of belief and disbelief are obtained as:

$$b_{c=c'}^{measure} = sim(c, c') \qquad d_{c=c'}^{measure} = 1 - b_{c=c'}^{measure}. \tag{4.2}$$

□

Corollary 4.2 (Discounting an opinion with a dogmatic opinion)**.** *Let A be a source who has an opinion about y in context c', expressed as*

$$\omega^A_{y:c'} = (b^A_{y:c'}, d^A_{y:c'}, u^A_{y:c'}, a^A_{y:c'})$$

and let the semantic similarity between the contexts c and c' be represented as a dogmatic opinion

$$\omega^{measure}_{c=c'} = (b^{measure}_{c=c'}, d^{measure}_{c=c'}, 0, a^{c'}_{c=c'}).$$

Since the source A does not have any prior opinion about the context c, we derive the opinion of A about c represented as

$$\omega^{A:c'}_c = (b^{A:c'}_c, d^{A:c'}_c, u^{A:c'}_c, a^{A:c'}_c)$$

using the base rate discounting operator on the dogmatic opinion.

$$\begin{aligned} a^{A:B}_y = a^B_y \qquad b^{A:B}_y = sim(c,c') \cdot b^B_y \\ u^{A:B}_y = 1 - sim(c,c') \cdot (b^B_y + d^B_y) \qquad d^{A:B}_y = sim(c,c') \cdot d^B_y. \end{aligned} \tag{4.3}$$

Definition 4.3 (Weighing operator)**.** Let C be the set of contexts c' of which a source A has an opinion derived from the positive and negative evidence in the past. Let c be a new context for which A has no opinion yet. We can derive the opinion of A about facts in c, by weighing the relevant evidences in set C with the semantic similarity measure $sim(c,c')$ $\forall c' \in C$. The belief, disbelief, uncertainty and a priori obtained through the weighing operation are expressed below.

$$\begin{aligned} b^A_c = \frac{sim(c,c') \cdot p^A_{c'}}{sim(c,c')(p^A_{c'} + n^A_{c'}) + 2} \qquad d^A_c = \frac{sim(c,c') \cdot n^A_{c'}}{sim(c,c')(p^A_{c'} + n^A_{c'}) + 2} \\ u^A_c = 1 - \frac{sim(c,c') \cdot (p^A_{c'} + n^A_{c'})}{sim(c,c')(p^A_{c'} + n^A_{c'}) + 2} \qquad a^A_c = a^A_{c'}. \end{aligned} \tag{4.4}$$

Theorem 4.4 (Approximation of the weighing and discounting operators)**.** *Let*

$$\omega^{A:c'}_{y:c} = (b^{A:c'}_{y:c}, d^{A:c'}_{y:c}, u^{A:c'}_{y:c}, a^{A:c'}_{y:c})$$

be a discounted opinion which source A has about y in a new or unknown context c, derived by discounting A's opinion on known contexts $c' \in C$ represented as $\omega^A_{c'} = (b^A_{c'}, d^A_{c'}, u^A_{c'}, a^A_{c'})$ with the corresponding dogmatic opinions (e.g. sim(c,c')). Let source A also obtain an opinion about the unknown context c based on the evidence available from the earlier contexts c', by weighing the evidence (positive and negative) with the semantic similarity

between c and c', $sim(c, c')$ $\forall c' \in C$. Then the difference between the results from the weighing and from the discount operator in subjective logic are statistically insignificant.

Proof. We substitute the values of belief, disbelief, uncertainty values in Equation (4.5) for Base Rate Discounting with the values from Equation (2.3)) and the expectation value from Equation (4.1). We obtain the new value of the discounted base rate opinion as follows:

$$\begin{aligned} b_c^{A:c'} &= \frac{sim(c,c') \cdot p_{c'}^A}{(p_{c'}^A + n_{c'}^A + 2)} & d_c^{A:c'} &= \frac{sim(c,c') \cdot n_{c'}^A}{(p_{c'}^A + n_{c'}^A + 2)} \\ u_c^{A:c'} &= 1 - \frac{sim(c,c') \cdot (p_{c'}^A + n_{c'}^A)}{(p_{c'}^A + n_{c'}^A + 2)} & a_c^{A:c'} &= a_{c'}^A. \end{aligned} \tag{4.5}$$

Equation (4.5) and (4.4) are pretty similar, except for the $sim(c, c') \cdot (p_{c'}^A + n_{c'}^A)$ factor in the weighing operator. In the following section we use a 95% student t-test and Wilcoxon signed-rank statistical test which shows that for our evaluation datasets the difference due to $sim(c, c') \cdot (p_{c'}^A + n_{c'}^A)$ factor is not statistically significant for large values of $sim(c, c')$. □

4.3.2 Evaluations

We show empirically the similarity between weighing and discounting.[1]

4.3.2.1 First experiment: discounting and weighing in a real-life case

We propose here a first validation of the similarity between weighing and discounting by using both of them in the process of estimation of the trustworthiness of a series of tags derived from `Steve.Museum` dataset.

Gathering evidence for evaluation We select a very small set of semantically related tags, by using a Web-based *WordNet* interface[2]. We then gather the list of users who provided the tags regarding the chosen words and count the number of positive and the negative evidence. The chosen tags are only three (Asian, Chinese and Buddhist), and they correspond to 206 entries in total (i.e., they are associated 206 times to one or more pictures by one or more users). This represents a small sample compared to the total number of tag entries (0.5%). However, this experiment is meant only to exemplify the use of the semantic similarity measure when one needs to compute an opinion about a new context (e.g., "Chinese"), given two existing ones (e.g., "Asian" and "Buddhist"). Therefore, we consider the *Chinese-Asian*

[1] Complete results at `https://github.com/anottamkandath/Datasets/tree/master/Chapter4`.
[2] `https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`

pair (semantic similarity 0.933) and the *Chinese-Buddhist* pair (semantic similarity 0.6667). We refer to the second experiment for a more indicative evaluation.

The opinions are calculated using two different methods. First by weighing the evidence with the semantic relatedness using Equation (4.4) and the second method is by discounting the evidence with the semantic relatedness using Equation (4.5).

Results We employ the Student t-test and the Wilcoxon signed-rank test to assess the statistical significance of the difference between two sample means. At 95% confidence level, both tests show a statistically significant difference between the two means. This difference, for the *Chinese-Asian* pair is 0.025, while for the *Chinese-Buddhist* pair is 0.11, thanks also to the high similarity (higher than 0.5) between the considered topics. Having removed the average difference from the results obtained from discounting (which, on average, are higher than those from weighing), both tests assure that the results of the two methods distribute equally.

4.3.2.2 Second experiment: discounting and weighing on a large simulated dataset

In the `Steve.Museum` dataset, the average number of annotations provided by a given user is limited (to about 20). To check if the two methods for building subjective opinions using semantic similarity measures are significantly different, we build a large dataset consisting of 1000 sample tags and treated the tags as if they were contributed by the same user. In this manner, we can check if the two methods present relevant differences both when the evidence amount is small or large. We perform the Student t-test and the Wilcoxon signed-rank test to evaluate the hypothesis that the two methods are not statistically significantly different. The results of the test show that for semantic relatedness values $sim(c, c') > 0.7$, the mean difference between the belief values obtained by weighing and discounting is 0.092. Thus with 95% confidence interval, both tests assure that both the weighing operator and the discounting operator produce similar results. The semantic similarity threshold $sim(c, c') > 0.7$ is relevant and reasonable, because it becomes more meaningful to compute opinions for a new context based on the opinions provided earlier for the most semantically related contexts, while also in case of lack of evidence for a given context, evidence about a very diverse context can not be very significant.

4.4 Combining probabilistic semantic similarity measures within subjective logic

The second extension that we propose regards the use of probabilistic semantic similarity measures within subjective logic.

4.4.1 Wikipedia relatedness measure

The Wu & Palmer measure described in Chapter 2 is a deterministic semantic similarity measure, because it is deterministically computed based on the position of the two examined words in *WordNet*. We propose the adoption in subjective logic of semantic similarity measures belonging to another class, that is, the probabilistic class of measures. These measures determine the semantic similarity between two words in a statistical manner, by checking the occurrence and co-occurrence of the two words within a large corpora of documents. A famous example of this kind of similarity measures is the Normalized Google Distance [21], which uses Google as a corpus of documents.

We use the Wikipedia[3] relatedness measure, as defined by Milne et al. [73, 74], because of its ease of use. This distance adapts the Normalized Google Distance by using Wikipedia as a corpus of reference for computation. The Wikipedia similarity distance is defined as follows:

$$sim(c, c') = \frac{log(max(|A|, |B|) - log(|AB|)}{log(|W|) - log(min(|A|, |B|)} \tag{4.6}$$

where $sim(c, c')$ is the semantic similarity between annotations c and c' and $|A|$ and $|B|$ are the cardinalities of the sets of Wikipedia documents containing c and c' respectively, and $|W|$ is the size of Wikipedia..

Milne et al. provide a disambiguation confidence score for the measure, that ranges between zero and one which is good as Subjective logic also computes trust score in the same range, thereby helping to combine Subjective logic techniques with Wikipedia relatedness measure.

[3] `http://www.wikipedia.org`

4.4.2 Wikipedia relatedness measure as a subjective opinion

As we discussed in Chapter 2, given two synsets (s_1 and s_2), we name c and c' the respective context identified by them. To differentiate from the previous section, we use $measure'$ as a placeholder for probabilistic similarity measures.

The elements at our disposal from the Wikipedia distance are:

- $sim(c, c') \in [0, 1]$ is the semantic relatedness between two annotations c and c';
- $conf(c, c') \in [0, 1]$ is the confidence in the semantic relatedness between c and c'.

To represent the Wikipedia distance in subjective logic, we need to map all its elements to specific elements (or combinations of elements) of subjective logic, while taking into account the logic's constraints and mechanisms (e.g., the fact that $b + d + u = 1$). We provide a mapping for each of the elements above, and we provide a motivation for them as follows.

1. $conf(c, c') = 1 - u_{c=c'}^{measure'}$ because the confidence value determines exactly the portion of probability mass that is certain. Therefore, the remaining part of the probability mass is assigned to the uncertainty element of subjective opinions.
2. $E_{c=c'}^{measure'} = sim(c, c')$. That is, the expected value of the subjective opinion should coincide with the similarity between the two annotations considered.
3. $b_{c=c'}^{measure'} = conf(c, c') \cdot sim(c, c')$ because the certain part of an opinion $(1 - u)$ is assigned $b+d$. Thus, we assign this mass proportionally to the value of the similarity measure, to represent our belief in the two annotations being semantically related.

However, given the constraints of subjective logic, by virtue of Equation (2.4) that we report as follows,

$$E_{c=c'}^{measure'} = b_{c=c'}^{measure'} + a_{c=c'}^{measure'} \cdot u_{c=c'}^{measure'}$$

we obtain

$$sim(c, c') = a_{c=c'}^{measure'}$$

which is, of course, wrong. The similarity value might depend on the subjective opinion's prior, but if the equation above holds, then we do not even need to compute the opinion, since the a priori value would already give the similarity value.

We propose two mappings between subjective opinions and probabilistic semantic similarity measures, each of them satisfying two of the three requirements above. Of the three

requirements, only the first one is considered as unavoidable, because of the definition of the uncertainty of subjective opinions.

Definition 4.5 (Wikipedia relatedness measure of two annotations as a subjective opinion (expected value as semantic similarity))**.** We define a subjective opinion capturing the similarity between c and c' using the Wikipedia distance as follows:

$$sim(c, c') \equiv \omega_{c=c'}^{mesaure'}(b_{c=c'}^{measure'}, d_{c=c'}^{measure'}, u_{c=c'}^{measure'}) \tag{4.7}$$

where

$$\begin{aligned} b_{c=c'}^{measure'} &= sim(c, c') - a_{c=c'}^{measure'} + a_{c=c'}^{measure'} \cdot conf(c, c') \\ d_{c=c'}^{measure'} &= sim(c, c') + a_{c=c'}^{measure'} - a_{c=c'}^{measure'} \cdot conf(c, c') + conf(c, c') \\ u_{c=c'}^{measure'} &= 1 - conf(c, c'), \end{aligned} \tag{4.8}$$

Using the values of b, u and a to compute E we get,

$$E_{c=c'}^{measure'} \equiv sim(c, c'). \tag{4.9}$$

We provide here motivation for the mapping that we propose. We treat the confidence value $conf(c, c')$ as the inverse of the uncertainty of a subjective opinion. In fact, we interpret the confidence as the percentage of probability mass confidently assigned by the semantic relatedness: the semantic relatedness ranges between zero and one, but we are confident on only $conf(c, c')\%$ of that mass. The rest of the probability mass $(1 - conf(c, c'))$ is, indeed, uncertain.

We also set the expected value of the opinion to coincide with the similarity value, that is:

$$E_{c=c'}^{measure'} = sim(c, c')$$

From this, given Equation (2.4), and having set $u_{c=c'}^{measure'} = 1 - conf(c, c')$, it follows that:

$$\begin{aligned} b_{c=c'}^{measure'} &= E_{c=c'}^{measure'} - a_{c=c'}^{measure'} \cdot (1 - conf(c, c')) \\ &= sim(c, c') - a_{c=c'}^{measure'} \cdot (1 - conf(c, c')) \\ &= sim(c, c') - a_{c=c'}^{measure'} + a_{c=c'}^{measure'} \cdot conf(c, c') \end{aligned}$$

and

$$\begin{aligned} d_{c=c'}^{measure'} &= 1 - b_{c=c'}^{measure'} - u_{c=c'}^{measure'} \\ &= 1 - (sim(c, c') - a_{c=c'}^{measure'} + a_{c=c'}^{measure'} \cdot conf(c, c')) - (1 - conf(c, c')), \end{aligned}$$

so

$$d_{c=c'}^{measure'} = sim(c, c') + a_{c=c'}^{measure'} - a_{c=c'}^{measure'} \cdot conf(c, c') + conf(c, c').$$

In this manner we define an opinion that reflects our constraints, that is: (1) uncertainty as inverse of the confidence of the semantic similarity value and (2) semantic similarity value as expected value of the subjective opinion. However, this mapping has the undesirable consequence that the belief $b_{c=c'}^{measure'}$ and the disbelief $d_{c=c'}^{measure'}$ depend on the a priori value $a_{c=c'}^{measure'}$. So, we propose an alternative mapping.

Definition 4.6 (Wikipedia relatedness measure of two annotations as a subjective opinion (belief as semantic similarity times confidence))**.** We propose here an alternative mapping that allows a subjective opinion to capture the similarity between annotations c and c' using the Wikipedia distance. The mapping is defined as follows:

$$sim(c, c') \equiv \omega_{c=c'}^{measure'}(b_{c=c'}^{measure'}, d_{c=c'}^{measure'}, u_{c=c'}^{measure'}), \tag{4.10}$$

where

$$\begin{aligned} b_{c=c'}^{measure'} &= conf(c, c') \cdot sim(c, c') \\ d_{c=c'}^{measure'} &= conf(c, c') \cdot (1 - sim(c, c')) \\ u_{c=c'}^{measure'} &= 1 - conf(c, c'). \end{aligned} \tag{4.11}$$

Again, we set the constraint $u_{c=c'}^{measure'} = 1 - conf(c, c')$, however we do not bind the expected value of the opinion to be equal to $sim(c, c')$.

We have shown that we can represent Wikipedia relatedness measures between two annotations by means of subjective opinions. As with many other subjective logic operators [58], we propose two possible mappings for the probabilistic semantic similarity measure. In particular, the second mapping that we propose does not present the undesirable characteristic shown by the first one, that is, a dependency between a priori value and belief in the mapped opinion. Of course, these two mappings are different, so we do not check their equivalence, like we did in the previous section for the mapping between subjective logic and probabilistic semantic similarity measures.

Our goal is to show how to represent semantic similarity measures in subjective logic, to import externally defined elements in the logic and increase its capabilities. The choice of the mapping is dependent on the specific constraints given by a domain or an application where the logic is used in combination with the similarity measure, although our preference goes to the second mapping, because the first one presents an already mentioned undesirable dependency between belief and a priori value. The same reasoning applies to the choice of the semantic similarity measure to adopt and it is a domain- and

application-dependent choice. Each semantic similarity measure has specific limitations. Deterministic semantic similarity measures are useful when the vocabulary of annotations provided to the cultural heritage institutions by annotators from the Web is from a known set of words (such as in the case of `Steve.Museum` dataset) and these words are easily classifiable and organised in a hierarchical knowledge graph (e.g. *Wordnet*). Probabilistic semantic similarity measures are useful when the annotations are more often words from the Web or words which undergo constant transformation in their usage. Since the probabilistic measures are computed based on a corpus of documents, the documents can be added or deleted and this would affect the semantic similarity measure computed from them.

4.5 Partial evidence observations

The Web and the Semantic Web are pervaded with data that can be used as evidence for a given purpose, but that constitute partially positive/negative evidence for others. In the *Waisda?* tagging game described in Chapter 2, users challenge each other about video tagging. The more users insert the same tag about the same video within the same time frame, the more the tag is believed to be correct. Matching tags can be seen as positive observations for a specific tag to be correct. However, consider the orthogonal issue of user reputation. It is based on past behaviour, hence on the trustworthiness of the tags previously inserted by him/her. Now, the trustworthiness of each tag is not deterministically computed, since it is roughly estimated from the number of matching tags for each tag inserted by the user. The expected value of each tag, which is at most one, can be considered as a partial observation of the trustworthiness of the tag itself. Vice versa, the remainder can be seen as a negative partial observation. After having considered tag trustworthiness, one can use each evaluation as partial evidence with respect to the user reliability: no tag (or other kind of observation) is used as a fully positive or fully negative evidence, unless its correctness has been proven by an authority or by another source of validation. However, since only rarely the belief (and therefore, the expected value) is equal to one, these observations almost never count as a fully positive or fully negative evidence. We propose an operator for building opinions based on indirect observations, i.e., on observations used to build these opinions, each of which counts as an evidence.

Theorem 4.7 (Partial evidence-based opinions)**.** *Let $\overrightarrow{P}$ be a vector of positive observations (e.g. a list of "hits" or "match" counts) about distinct facts related to a given subject s. Let l be the length of $\overrightarrow{P}$. Let each opinion based on each entry of $\overrightarrow{P}$ have an a priori value of $\frac{1}{2}$. Then we can derive an opinion about the reliability of the subject in one of the following two manners.*

- *By cumulating the expected values (counted as partial positive evidence) of each opinion based on each element of p:*

$$b_s = \frac{1}{l+2} \cdot \sum_{i=1}^{l} \frac{p_i+1}{p_i+2} \quad d_s = \frac{1}{l+2} \cdot \sum_{i=1}^{l} \frac{1}{p_i+2} \quad u_s = \frac{2}{l+2}. \tag{4.12}$$

- *By averaging the expected values of the opinions computed on each of the elements of p:*

$$b_s = \frac{1}{3l} \cdot \sum_{i=1}^{l} \frac{p_i+1}{p_i+2} \quad d_s = \frac{1}{3} - \frac{1}{3l} \cdot \sum_{i=1}^{l} \frac{1}{p_i+2} \quad u_s = \frac{2}{3}. \tag{4.13}$$

Proof. For each "fact" about s we have at our disposal a count of positive pieces of evidence. We treat each fact as an observation about the trustworthiness of s. Examples of these observations are tags inserted by s in a crowdsourcing platform. The items of evidence are the approvals or matches that these tags obtain. We do not set an upper limit to the amount of positive evidence. Rather, we convert it into a subjective opinion and we compute its expected value as follows (remember that no negative evidence is registered):

$$E_i = b_i + a_i \cdot u = \frac{p_i}{p_i+2} + \frac{1}{2} \cdot \frac{2}{p_i+2} = \frac{p_i+1}{p_i+2}. \tag{4.14}$$

E is considered as partial positive evidence. If p is an extremely high number, then E is approximated to 1. Otherwise, $1-E$ is considered as partial negative evidence. Given that we have l pieces of partial evidence (because we have l distinct elements in $\overrightarrow{P}$), we compute the opinion about s following Equations (2.3). Here we have two possibilities. If l contains evidence about distinct and independent facts, then we can cumulate all the pieces of evidence (represented as $E_i, 1-E_i$), and by setting:

$$p_s = \sum_{i=1}^{l} \frac{p_i+1}{p_i+2} \qquad n_s = \sum_{i=1}^{l} \frac{1}{p_i+2},$$

we obtain Equation (4.12). In fact, we consider each item of $\overrightarrow{P}$ as providing an observation about s.

If, instead, $\overrightarrow{P}$ contains dependent observations, then it makes sense to average them in order to uniformly represent the evidence about s. In this case, we set:

$$p_s = \frac{1}{l} \cdot \sum_{i=1}^{l} \frac{p_i+1}{p_i+2} \qquad n_s = \frac{1}{l} \cdot \sum_{i=1}^{l} \frac{1}{p_i+2}.$$

Following again Equations (2.3), we obtain Equation (4.13). Note that in this case we use only the average of the observation as item of evidence. Therefore, we have only

one item of evidence. This justifies the fact that in Equation (4.13) we always have 3 as denominator: following Equation (2.3), $p + n + 2 = 1 + 2 = 3$. □

For the example below, we use Equation (4.12), because we consider the cases where evidence from different and independent facts about the same individual are provided.

Suppose in *Waisda?* tagging game, a user *user* added two different tags about two different videos. One of them got five matches, while the other got two. We can compute a subjective opinion about *user* that represents his reputation using Equation (4.12) and we obtain:

$$\omega_{user}\left(\frac{1}{l+2}\cdot\sum_{i=1}^{l}\frac{p_i+1}{p_i+2}, \frac{1}{l+2}\cdot\sum_{i=1}^{l}\frac{1}{p_i+2}, u_s = \frac{2}{l+2}\right) =$$
$$\omega_{user}\left(\frac{1}{4}\left(\frac{6}{7}+\frac{3}{4}\right), \frac{1}{4}\left(\frac{1}{7}+\frac{1}{4}\right), \frac{1}{4}\left(\frac{2}{4}\right)\right) =$$
$$\omega_{user}\left(\frac{55}{112}, \frac{11}{112}, \frac{1}{2}\right).$$

If, instead, the two tags got the same scores as before, but they were inserted for the same video in different matches, we can average their contributions, since they provide indications about the user reliability in the same situation. What we obtain using Equation (4.13) is:

$$\omega_{user}\left(\frac{1}{3l}\cdot\sum_{i=1}^{l}\frac{p_i+1}{p_i+2}, \frac{1}{3} - \frac{1}{3l}\cdot\sum_{i=1}^{l}\frac{1}{p_i+2}, \frac{2}{3}\right) =$$
$$\omega_{user}\left(\frac{1}{6}\left(\frac{6}{7}+\frac{3}{4}\right), \frac{1}{3} - \frac{1}{6}\left(\frac{6}{7}+\frac{3}{4}\right), \frac{2}{3}\right) =$$
$$\omega_{user}\left(\frac{55}{168}, \frac{1}{168}, \frac{2}{3}\right).$$

4.6 Conclusion

We have shown the potential for employing subjective logic as a basis for reasoning on Web and Semantic Web data. We showed that it can be powerful for handling uncertainty and how little extensions can help in improving the mutual benefit that Semantic Web and subjective logic obtain from cooperating together. We proposed the use of semantic similarity measures, both deterministic (in particular, the Wu & Palmer similarity measure) and probabilistic ones (in particular, the Wikipedia semantic relatedness), within subjective logic.

Furthermore, we proposed a means to represent subjective opinions on the basis of partial evidence, which is a common phenomenon on the Web (e.g. number of hits or number of tweets). This operator will be employed for experiments and evaluation in Chapter 5.

Chapter 5

Combining Reputation-based and Provenance-based Trust

In this chapter trust concerns are handled by leveraging reputation of annotators and provenance of annotation process. The work in this chapter was first presented at the 8th International Workshop on Uncertainty Reasoning for the Semantic Web at the 11th International Semantic Web Conference (ISWC 2012) in Boston, USA and later published in Uncertainty Reasoning for the Semantic Web III - Revised Selected Papers of URSW 2011-2013, Lecture Notes in Artificial Intelligence 8816, Springer LNAI 8816 Proceedings in 2014. My contributions are in the design, implementation and evaluation for the section Analysis of correlation between user demographics and data trustworthiness and in the section Computing reputation-based trust.

5.1 Introduction

In this chapter, firstly we perform a series of analyses to demonstrate the existence of correlations between user demographics and the trustworthiness of the data they provide. Secondly, we compute reputation of annotators by using subjective logic and weighing on time factor instead of semantic similarity measures as done in the previous chapters. Thirdly, we propose a procedure for computing trust assessments based on provenance information represented in the W3C PROV model. Such a procedure is important because it is not always possible to have complete user demographic information. Here PROV plays a key role because of its ability to provide an interchangeable format: having modelled our procedure on PROV, any other different input format can be easily treated after having been mapped to PROV. We implement this procedure by discretising the trust values and applying a support vector machine (SVM) classification. Finally,

we combine these two procedures in order to maximise the benefit of both. The procedures are evaluated on data provided by the *Waisda?* dataset where users challenge each other in tagging videos. We show how to use the FOAF ontology to represent the user information provided in their profiles, and we provide a small extension of it to represent user stereotypes. A stereotype is an abstraction of user demographics. We then provide a procedure to compute the user trustworthiness based on stereotypes from information in user profiles. Through our experiments, we try to determine correlations between the trust of the users and the stereotype of their profile.

We show that a reputation-based prediction is not significantly different from a provenance-based prediction and, by combining the two, we obtain a small but statistically significant improvement in our predictions. We also show that reputation-based and provenance-based assessments correlate and that there is a correlation between the user profile stereotypes and the trust in a user.

5.2 Related work

The first part of our work focuses on reputation estimation and is inspired by the works collected by Masum and Tovey [69]. Pantola et al. [79] present reputation systems that measure the overall reputation of the authors based on the quality of their contribution and the "seriousness" of their ratings; Javanmardi et al. [51] measure reputation based on user edit patterns and statistics. Their approaches are similar to ours, but they are particularly tailored to wiki-based environments. The second part of our work focuses on the usage of provenance information for estimating trust assessments. In their works, Bizer and Cyganiak [8], Hartig and Zhao [42] and Zaihrayeu et al. [108], use provenance and background information expressed as annotated or named graphs [13] to produce trust values. We do not make use of annotated or named graphs, but we use provenance graphs as features for classifying the trustworthiness of artefacts. The same difference also applies to the two works of Rajbhandari et al. [82, 83], where they quantify the trustworthiness of scientific workflows and they evaluate it by means of probabilistic and fuzzy models. Provenance is used for data verification in crowdsourced environments by Ebden et al. [28]. In their work, they introduced provenance tracking into their online CollabMap application (used to crowdsource evacuation maps), and in this way they collect approximately 5,000 provenance graphs, generated using the Open Provenance Model [75] (which has now been superceded by the PROV Data Model and Ontology). In their work they have at their disposal large provenance graphs and can learn useful features about the artefact trustworthiness from the graphs topologies. Here, the graphs at our disposal are much more limited, so we cannot rely on the graph topology, but we

can easily group graphs in stereotypes. Provenance mechanisms have also been used to understand and study workflows in collaborative environments as discussed in Altintas et al. [2]. We share the same context with that work, but we do not focus on the workflow of artefact creation.

In the current chapter, we represent trust values by means of subjective opinions, but trust assessments are made by means of support vector machines, eventually combined with reputations, again represented by means of subjective opinions. The impact of user information such as age, gender, education and demographics in crowd sourcing tasks have been explored in the works of Kazai et al. [61] which explores the relationship between worker characteristics and the quality of their work. That work has been applied to the crowdsourcing domain and has proven that both the demographics and personality profiles of the workers are strongly linked to the resulting label quality. We apply our algorithm not on a labelling task on a crowdsourcing platform, but on a video annotation task.

Another work by Venanzi et al. [99] addresses the issue of having too few labels from a user to determine their quality by using a community-based Bayesian label aggregation model which assumes that crowd workers conform to a few different types, where each type represents a group of workers with similar confusion matrices. We use a similar approach to build stereotypes of users behaviour based on information provided by the users, but not for crowdsourcing systems. Their work is performed on the labeling task while ours is done on annotations of videos. In general, much work has been done in crowdsourcing platforms to determine the effect of a user profile on user accuracy and reputation (see [61], [99]). However, these works focus mainly on labeling crowdsourced data where ground truth data is already available. The main difference between our work on determining correlation of user profiles on their quality with the above mentioned work is that we do not have a ground truth. For the labelling tasks on the crowdsourcing platforms, there is ground truth available for both works. In our case, we lack such information and thus rely on partial evidence, which is that we trust a tag provided by a user more if there are other users who provided the same tag into the system. The procedure introduced in Section 5.5 is a generalisation of the procedure that we introduced in Chapter 3 where we evaluated the trustworthiness of tags of the `Steve.Museum` dataset.

Lastly, the use of stereotyping as a bootstrapping method has already been investigated by Liu et al. [66] and Burnett et al. [11]. There exist relevant similarities between these works and ours, like, for example, the use of subjective logic to represent trust (this probabilistic logic makes use of Beta and Dirichlet distributions to model trust statistically) and the fact that users can be grouped in stereotypes to obtain useful

information to assess unknown users. Nevertheless, there are also relevant differences. In fact, both these papers take an agent-based approach and their final goal is to determine whether we can trust an agent or not. Our goal, instead, is to determine the agent's (user's) trustworthiness to be able to use it to determine the trustworthiness of the artefact that he or she produces. Also, Burnett et al. proposed that agents can learn a stereotyping function, and Liu et al. propose that stereotyping is based on a function, although they do not investigate it further. In our work, we propose to create stereotypes based on user characteristics (and hence, implicitly, on a function of these characteristics), although we do not explicitly characterize this function.

5.3 Dataset processing

In this chapter, we use the *Waisda?* dataset to perform our experiments and the corpus contains 37,850 annotations. In Section 5.4, in order to determine the correlation of user profile information with user reputation, we used the data from 17 users who provided information about themselves in their user profiles. The remaining users did not provide their data or chose to remain anonymous. Initially, we tried to cluster the users based on their features such as age, number of contributions etc., and tried to draw conclusions about certain stereotypes. However, since we had too few users to draw conclusions based on this approach, we opted, instead, to use standard correlation metrics on our data. We used the Pearson correlation for the continuous data such as the number of tags provided, the number of tags provided which were matched with others, age, etc.. For categorical variables such as gender, we used the point biserial correlation metric.

For sections 5.4, 5.6 and 5.7, where we compute trust based on reputation, provenance and a combination of both, we used split the annotations into training and test set. We used 26,495 tag entries (70%) as a training set, and the remaining 11,355 (30%) as a test set.

We also were interested to see if the characteristics of the entire dataset are also observed in smaller samples of the dataset. For this purpose, we performed a random sampling of 115 annotations, which correspond to about 9% of the total population and compared the characteristics of this sample with that of the entire dataset. First, we compared the distribution of each relevant feature that we will use in Section 5.6 in our sample with the distribution of the same feature in the entire dataset. A 95% confidence level Chi-squared test confirmed that the hour of the day and the day of the week distribute similarly in our sample and in the entire dataset. The typing duration distributions (i.e., distributions of the time employed by users to insert annotations) instead, are significantly different according to a 95% confidence level Wilcoxon signed-rank test. However, the mode of the

two distributions are the same, and the mean differs only 0.1 seconds which, according to the Keystroke-level Model-GOMS model (KLM-GOMS) [12], corresponds, at most, to a keystroke. The KLM-GOMS predicts how long it will take an expert user to accomplish a routine task without errors using an interactive computer system. So we can conclude that the selected sample is representative for the entire data set. A second analysis showed that, by randomly selecting another sets of 115 annotations, the corresponding characteristics are not statistically different from the sample that we selected.

5.4 Analysis of correlation between user demographics and data trustworthiness

Demographics provide a set of quantifiable statistics about a population. A user profile is a collection of personal information about a given user. In this work, we assume that information taken in the aggregate from user profiles represents the demographics of the population.

Here, we try to determine if there is a correlation between the user reputation and demographics in the *Waisda?* dataset. We use the user reputation as a proxy for data trustworthiness.

Our analysis is performed by grouping users based on their demographics and by identifying a correlation between user groups and the trustworthiness of the artefacts they produced[1]. The drawback of our approach is that the users need to provide their details to the system. Since *Waisda?* is an online game, many users chose to participate as anonymous. We realised that the users who actively returned back to the game are mostly the ones who provided their profile information. This is a good indication of which users will actively participate in the system for a longer time. Another thing to note is that, in general, users may not provide accurate information about themselves in their profile. However, for the sake of this work, we do not take this possibility into account, because the users that provided their personal information in the game are known, and hence their information is trusted. Moreover, information inaccuracies (which can be incorrect data in the profile entered by the users), if any, are compensated since we take a statistical approach. The reason why we investigate the correlation between demographics and data trustworthiness is that we hypothesise that certain categories of users may be performing better than others. For instance, younger users may be more attentive while older users may be more accurate. If that is the case, then the stereotype

[1]Complete results at `https://github.com/anottamkandath/Datasets/tree/master/Chapter5`

that we define should help us in identifying groups of users whose performance is higher or lower than others.

5.4.1 User profiles and their representation

The information in the user profile and other quantitative information derived about a user can help to estimate user reputation. Although different systems gather different types of information from a user, there is an overlap between the most common features such as the age, gender, education, etc. Such information provided by the user can be represented using the FOAF ontology. FOAF provides a representation of the individual user along with his details. Apart from the user-provided details, we also derive information such as the number of tags contributed by the user, the percentage of tags matched with other users, etc. For representing data that are specific to the tagging environment and system, we do not adopt a standard and use an ad-hoc representation (the property *ex:contributed_tags* for the number of user contributed tags, and the property *ex:matched_tags* for the number of matched tags for a given user).

In our procedure, we also build groups (or stereotypes) of users who share similar characteristics. In order to form groups of users, we use percentiles for each characteristic in their profile and derived characteristics. Percentiles help in obtaining an even distribution of the users across different profile characteristics and grouping them in stereotypes. One example of a stereotype can be users who are at least thirty years old and female. In order to represent these groups or stereotypes, we utilize the *group* class of FOAF. The groups are formed based on the information in the individual FOAF profile. Figure 5.1 depicts an example of users Alice and Mary who are both females above thirty years of age and belong to the same Foaf:Group. We propose to use Foaf:Group class with Foaf:Person sub-classes to represent stereotypes. The fact that we use FOAF and a small extension of it is important, because it eases interoperability with the systems that use this widely adopted ontology.

In the next section, we explain a procedure for predicting the reputation of a user based on the aggregation of the reputations of users within the same stereotype.

5.4.2 Procedure for analyzing the correlation between user demographics and reputation

In order to evaluate the correlation between user demographics and the trustworthiness of the artefacts that they produce, we developed a procedure that groups users in stereotypes according to their personal information, and we check the existence of correlations

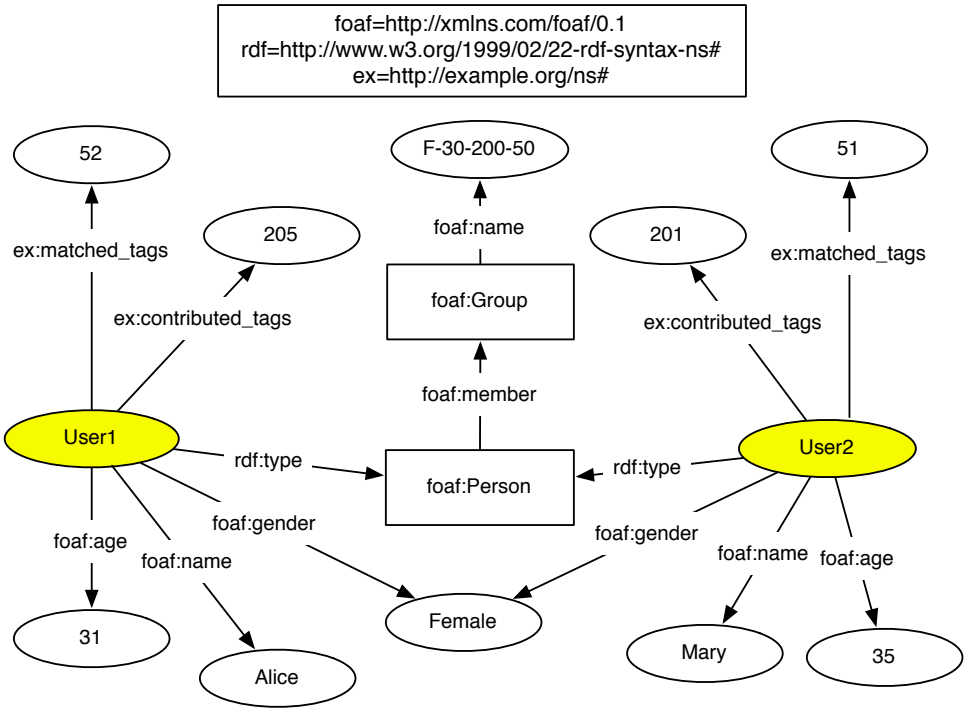

FIGURE 5.1: Graph representation of the users and groups. The group name F-30-200-50 is formed by female users that are older than thirty and provided more than 200 tags of which more than 50 are matched.

between the fact that a given user belongs to a certain stereotype and their reputation.

The procedure is as follows:

Algorithm 2: Procedure for making user profile based trust estimation.

```
1 Procedure reputation_profile_prediction(user, reputation, user_profile)
2     attribute_set ← attribute_selection(user_profile)
3     attributes ← attribute_extraction(attribute_profile)
4     trainingset, testset ← trust_levels_aggregation(trainingset,testset)
5     classified_testset ← classify(testset,trainingset)
6     return classified_testset
```

The subprocedures used are described below:

attribute_selection Among all the profile information provided by the user, the first step of our procedure chooses the most significant ones: age and gender. In this process we also distinguish between the categorical variables and the continuous variables. This selection can lead to an optimisation of the computation. As shown in Equation (5.1), the reputation of the user is influenced by the characteristics in his profile.

$$user\_reputation = age \otimes education \otimes gender \otimes salary \otimes \ldots \quad (5.1)$$

where $\otimes$ is a function which models the relationship between user_reputation and the different characteristics of the user profile.

attribute_extraction Apart from the user-provided information in the profile, we derive information about the user contributions in the system. This information can be the total number of tags provided, total number of tags matched with the other users, time spent in the system, etc. This derivation can help us understand the behaviour of the user better and help derive useful correlations between user behaviour and reputation.

trust_levels_aggregation To ease the learning process, we aggregate the reputations of the users into n classes. The classes are formed by different combinations of subclasses. The subclasses are created based on the extracted user information. To create a subclass, we compute percentiles for continuous variables such as age, total tags contributed and no: of matched tags. Using percentiles, we divide the continuous variables into four subclasses with each subclass containing a quarter of the data. For categorical variables such as gender and education we use each of the categories available. Once the classes are formed, we consider them as stereotypes of the users. We assign each user to a particular stereotype.

classify Machine learning algorithms (or any other kind of classification algorithm) can be adopted at this stage. The choice can be constrained either from the data or by other limitations (e.g., computational power at our disposal). In this subprocedure, we try to predict information about the reputation of a new user belonging to a certain stereotype based on the reputation of other users belonging to that particular stereotype. This prediction helps to give an "a priori value" of reputation for new users in the system based on information in their profiles.

5.4.3 Application evaluation

We apply the procedure to the tag entries from the *Waisda?* dataset as follows.

attribute selection and extraction In the *Waisda?* dataset, we have 17 users who provided in their profile personal information such as e-mail, id, age and gender. The remaining users of the tagging game participated as anonymous. We extract the age and gender from the profiles and derive information such as total number of tags contributed by each user and, for each user, the total number of tags matched with the others. We also compute the reputation of the users using the partial evidence extension of subjective logic that we introduced in Section 4.5 in Chapter 4.

trust based stereotypes computation We split the continuous variables such as age, total number of tags contributed, and total number of tags matched into four subclasses using percentiles. Each subclass has a quarter of the total data. We use this approach to ensure equal distribution of the data into the different subclasses. Categorical variables such as gender are divided into two subclasses (male, female). Once the subclasses are formed, we aggregate the subclasses in different combinations to form the stereotypes of the users. In our case for the *Waisda?* dataset we have seven stereotypes.

classify We used a regression algorithm to predict the trustworthiness of the users belonging to a stereotype. Once we have sufficient evidence (e.g. at least five or ten users belonging to a stereotype), we can predict the trustworthiness of new users in the system who belong to the same stereotype. This prediction can help us to give an idea about the user trustworthiness in the system and also in the future help to recruit users with certain characteristics for the system.

5.4.4 Results

Table 5.1 shows the results of our analysis about the user reputation per stereotype. Here the user reputation is computed by using the formulas presented in Section 4.5 and by executing steps listed in Subsection 5.5.2: for each user in each stereotype we compute the frequency of matched tags that he or she contributed, weighed on the sample size.

Table 5.1: Stereotypes of user profiles and their reputation

Stereotype	# users	User reputations
Stereotype 1	2	[0.96, 0.90]
Stereotype 2	2	[0.97, 0.95]
Stereotype 3	2	[0.91, 0.94]
Stereotype 4	2	[0.97, 0.96]
Stereotype 5	5	[0.97, 0.97, 0.97, 0.98, 0.98]
Stereotype 6	1	[0.94]
Stereotype 7	3	[0.95, 0.93, 0.95]

From Table 5.1, we observe that there is not much difference between the reputation values of the users belonging to the different stereotype groups. Also, the difference within stereotypes is very small. The results cannot help us make generalisations since the sample size of the data is very small. The reason why we presented the results is to show the applicability of our procedure on the dataset. In Chapter 7, we have a larger dataset and apply many statistical tests to observe the correlation between user profile metrics and reputation. The work in this chapter serves as an introduction to new techniques for determining quality of information based on user profile data.

Stereotypes based only on demographics features that correlate with user reputations may be able to discriminate users on their reputations. Also, in this specific use case, the variance of the user reputation is quite low, so it may be hard to group users based on their reputation. We also try to evaluate the correlation between user demographics and user reputation, since the sample size becomes 17, which is the number of users. decompose the information that determines the user stereotype and we analyse these components independently. For data which is normally distributed, we use the Pearson correlation. For categorical data such as age, we use point biserial correlation. The results of our analysis are shown in Table 5.2.

TABLE 5.2: Results of correlation analysis on *Waisda?* dataset

X	Y	Correlation method	Corr(X,Y)	p-value
# of tags	Reputation	Pearson	0.53	0.02
# of matched tags	Reputation	Pearson	0.61	0.008
Age	Reputation	Pearson	-0.55	0.02
Gender	Reputation	Point biserial	0.46	0.06

From Table 5.2 we can see that there is linear positive correlation between the percentage of tags provided by a user that match with other tags and the user's reputation. However, there is a negative correlation with the user age and their reputation. The point biserial correlation method shows that there is a positive correlation between the gender of users and their reputation.

Statistical correlation tests provide better results when the sample size is large. In our case we have a sample size of 17 and thus it is hard to make generalisations based on the results of the tests. However we report the observations. It can be seen that there is a correlation between the information provided by the user and their reputation, at least in the *Waisda?* dataset. For instance, the age correlation indicates that the youngest users perform best, perhaps because they are more reactive and attentive. Also, users that contributed more tags tend to have a higher reputation. This is probably because they developed a better tagging skill over time. Users that contributed a higher number of matched tags also tend to be more precise. (A higher number of matched tags does not need to correspond to a higher reputation, since the matched tags could be accompanied by a lot of unmatched ones; this is not the case here.) The gender correlation is not significant, since it is even lower than the probability to guess the correct reputation to a user based on his or her gender. These correlations can help us to predict the reputation of new users based on reputations computed from users with similar characteristics. The results from this study may be useful for expert finding, since once we learn which stereotypes of users perform a certain task well, we can recruit more users of that stereotype into the system.

5.5 Computing reputation-based trust

In the previous section, we analysed some of the assumptions that underpin the use of user reputations for making trust assessments. We find that there exists a moderate correlation between the user demographics and the trustworthiness of the data that the population produces. This leads us to conclude that by virtue of the correlation between user reputation and demographics, demographics can be used as a foundation for trust prediction, although particular countermeasures need to be taken to compensate for the fact that the existing correlation is only moderate.

Here, we provide a generic procedure that allows to build a reputation for a user, based on a set of evaluated artefacts (e.g., annotations), and to use it for assessing trust of other artefacts created by him. We build the reputation based on a set of evaluated tags contributed by the user and not on user demographics because we have such evaluations at our disposal, and this allows tailoring the reputation to the specific user. Still, the analysis presented before lays the foundations for the use of user reputation for trust prediction.

5.5.1 Procedure

We present a generic procedure for computing the reputation of a user with respect to a given artefact produced by him.

Algorithm 3: Procedure for reputation computation.

```
1 Procedure reputation(user,artefact)
2     evidence ← evidence_selection(user,artefact)
3     weighted ← evidence_weighing(user, artefact, evidence)
4     reputation ← aggregate_evidence(weighted_ evidence)
5     return reputation
```

Evidence_selection Reputation is based on historical evidence, hence the first step is to gather all pieces of evidence regarding a given user and select those relevant for trust computation. Typical constraints include temporal (evidence is only considered within a particular time-frame) or semantics (evidence is only considered when it is semantically related to the given artefact). We define *evidence* as the set of all evidence regarding *user* about *artefact*.

Evidence_weighing Given the set of evidence considered, we can decide if and how to weigh its elements, that is, whether to count all the pieces of evidence as equally important, or whether to consider some of them as more relevant. This step might

Algorithm 4: Procedure for evidence selection.

```
1 Procedure evidence_selection(user,artefact)
2     for i ← 1 to length(observations) do
3         if observations[i].user = user then
4             evidence.add(observation[i])
5         end
6
7     end
8     return evidence
```

be considered as overlapping with the previous one since they are both about weighing evidence: evidence selection gives a boolean weight, while here a fuzzy or probabilistic weight is given. However, keeping this division produces an efficiency gain, since it allows computation to be performed only on relevant items.

Algorithm 5: Procedure for weighing evidence.

```
1 Procedure evidence_weighing(user,artefact,evidence)
2     for i ← 1 to length(evidence) do
3         weighted_evidence.add(weigh(evidence[i],artefact))
4     end
5
6     return weighted_evidence
```

Aggregate_evidence Once the pieces of evidence have been selected and weighed, these are aggregated to provide a value for the user reputation that can be used for evaluation. We can apply several different aggregation functions, depending on the domain. Typical functions are: *count*, *sum*, *average*. We use subjective logic which is described in detail in Chapter 2 for the application of our procedure.

5.5.2 Application evaluation

First, we convert the number of matches that each tag entry has into trust values. We obtain an opinion for a given tag entry by aggregating all the evidence (in the form of match or non-match) from the other tag entries. For brevity, we report the details about the computation of p and n (i.e. of the positive and negative evidence counts). The corresponding subjective opinion is always computed as discussed in Chapter 2.

tag selection For each tag inserted by the user, we select all the matching tags belonging to the same video. In other contexts, the number of matching tags can be substituted by the number of "likes", "retweets", etc.

tag entries weighing For each matching entry, we weigh it on the time distance between the evaluated entry and the matched entry. The weight is determined from an exponential probability distribution, which is a "memory-less" probability distribution used to describe the time between events. If two entries are close in time, we consider it highly likely that they match. If they match but appear at distant temporal moments, then we presume they refer to different elements of the same video. Instead of choosing a threshold, we give a probabilistic weight to the matching entry. 85% of probability mass is assigned to tags inserted in a ten seconds range.

tag entries aggregation In this step, we determine the trustworthiness of every tag. We aggregate the weighed evidence in a subjective opinion about the tag trustworthiness. We have at our disposal only positive evidence (the number of matching entries). The more evidence we have at our disposal for the same tag entry, the less uncertain our estimate of its trustworthiness will be. Non-matched tag entries have equal probability to be correct or not. We repeat the procedure above for each tag entry created by the user to compute his reputation.

user tag entries selection Select all the tag entries inserted by *user*.

user tag entries weighing Tag entries are weighed by the corresponding trust value previously computed. If an entry is not matched, it is considered as a half positive (tag trust value 0.5) and half negative ($1 - 0.5 = 0.5$) item of evidence (it has 50% probability to be incorrect), as computed by means of subjective opinions. The other entries are also weighed according to their trust value. So, user reputation can either rise or decrease as we collect evidence. We compute trust values of tag entries using the technique for determining trust when partial evidence is available, which is discussed in Chapter 4.

user tag entries aggregation In turn, to compute the reputation of a user with respect to a given tag, we use all the previously computed evidence to build a subjective opinion about the user. This opinion represents the user reputation and can be summarized even more by the corresponding expected value.

5.5.3 Results

We implement the abstract procedure for reputation computation and we evaluate its performance by measuring its ability to make use of the available evidence to compute the best possible trust assessment. Our evaluation does not focus on the ability to predict the exact trust value of the artefact by computing the user reputation, because these two

values belong to a continuous space, and they are computed using different techniques. What we expect is that these two values hint at trustworthiness in a similar fashion. We suppose that the trust evaluation system is implemented in such a manner that tags are "accepted" as trustworthy when their trust value is higher than a particular value (also called threshold). So, if the user reputation is a good indicator of trustworthiness, the reputation of a user should be higher than the threshold when the trust values of the artefacts created by him pass the threshold, and vice-versa. The validation, then, depends upon the choice of the threshold which, in turn, depends on constraints imposed by each specific use case. For instance, as we explain below, in the case study we tackle, "false negatives" are preferred over "false positives", and this makes the threshold more likely to be set high (e.g., at least 85% or 90%).

We run the procedure with different thresholds as presented in Figure 5.3. Low thresholds correspond to a low accuracy in our predictions. However, as the threshold increases, the accuracy of the prediction rises. Moreover, we should consider that: (1) it is preferable to obtain "false negatives" (reject correct tags) rather than "false positives" (accept wrong tags), so high thresholds are more likely to be chosen (e.g., see [33]), in order to reduce risks. Rejecting correct tags means rejecting useful information and therefore wasting part of the effort spent in crowdsourcing tags. Accepting wrong tags means to introduce in the system wrong information and therefore, the tasks that rely on these crowdsourced tags may be affected by this (e.g. if we run an information retrieval task using these tags, then we may retrieve wrong items). Hence we prefer the first situation in place of the latter; (2) a Wilcoxon signed-rank test at 95% confidence level proved that the reputation-based estimates outperform blind guess estimates (having average probability of accuracy 50%). The average improvement is 8%, the maximum is 49%.

We previously adopted this procedure to compute the trustworthiness of tags on the `Steve.Museum` dataset in Chapter 3. By adapting the procedure to the *Waisda?* dataset, we were able to formulate the general procedure above.

5.6 Computing provenance-based trust

User demographics and, in general, user identities are not always available when estimating the trustworthiness of artefacts. Hence, we provide a procedure for estimating the trustworthiness of artefacts based on "how" they were produced rather then on "who" produced them. Thus, we focus on the "how" part of provenance, i.e., the steps or activities performed in the production of an artefact. (For simplicity, in the rest of the chapter, we will use the word "provenance" to refer to the "how" part.) We learn the relationships between PROV (described in detail in Chapter 2) and trust values through

machine learning algorithms. This procedure allows to process PROV data and, on the basis of previous trust evaluations, predict the trust level of artefacts.

5.6.1 Procedure

We present the procedure for computing trust estimates based on provenance.

Algorithm 6: Procedure for making provenance-based estimation.

```
1 Procedure provenance_estimation(artefact_provenance,artefact)
2 |  attribute_set ← attribute_selection(artefact_provenance)
3 |  attributes ← attribute_extraction(attribute_set)
4 |  trainingset, testset ← trust_levels_aggregation(trainingset,testset)
5 |  classified_testset ← classify(testset,trainingset)
6 |  return classified_testset
```

attribute_selection Among all the provenance information, the first step of our procedure chooses the most significant ones: agent, processes, temporal annotations and input artefacts can all hint at the trustworthiness of the output artefact. This selection can lead to an optimization of the computation.

attribute_extraction Some attributes need to be manipulated to be used for our classifications; e.g., temporal attributes may be useful for our estimates because one particular date may be particularly prolific for the trustworthiness of artefacts. However, to ease the recognition of patterns within these provenance data, we extract the day of the week or the hour of the day of production, rather than the precise timestamp. In this way we can distinguish, e.g., between day and night hours (when the user might be less reliable). Similarly, we might refer to process types or patterns instead of specific process instances.

trust_level_aggregation To ease the learning process, we aggregate trust levels in n classes. Our results will show that this classification process does not affect accuracy significantly.

classify Machine learning algorithms (or any other kind of classification algorithm) can be adopted at this stage. The choice can be constrained either from the data or by other limitations.

5.6.2 Application evaluation

We apply the procedure to the tag entries from the *Waisda?* dataset as follows.

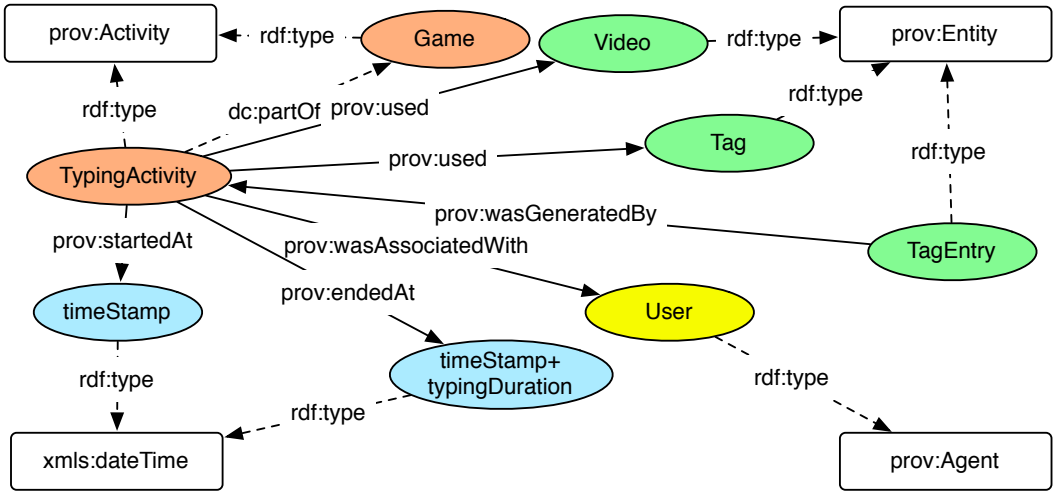

FIGURE 5.2: Graph representation of the provenance information about each tag entry.

attribute selection and extraction The provenance information available in *Waisda?* is represented in Figure 5.2, using the W3C PROV ontology. First, for each tag entry we extract: *typing duration*, *day of the week*, *hour of the day*, *game_id (to which the tag entry belongs)*, *video_id*. This is the "how" provenance information at our disposal. Here we want to determine the trustworthiness of a tag given the modality with which it was produced, rather than the author reputation. Some videos may be easier to annotate than others, or, as we mentioned earlier, user reliability can decrease during the night. For similar reasons we use all the other available features.

trust level aggregation In our procedure, we are not interested in predicting the exact trust value of a tag entry. Rather we want to predict the range of trust values that hold for an entry. Given the range of trust values $[0, 1]$, we split it into 20 classes of length 0.5: from $[0, 0.05]$ to $[0.95, 1.0]$. This allows us to increase the accuracy of our classification algorithm without compromising the accuracy of the predicted value or the computation cost. The values in each class were approximated by the middle value of the class itself. For instance, the class $[0.5 \ldots 0.55]$ is approximated as 0.525.

classify We use a regression algorithm to predict the trustworthiness of the tags. Having at our disposal five different features (in principle, we might have more), and given that we are not interested in predicting the "right" trust value, but the class of trustworthiness, we adopt the "regression-by-discretization" approach which is described in Chapter 2. The training set is composed by 70% of our data, and then we predict the trust level of the test set. We used the SVM version implemented in the e1071 R library.

5.6.3 Results

The accuracy of our predictions depends, again, on the choice of a threshold. If we look at the ability to predict the right (class of) trust values, then the accuracy is about 32% (which still is twice as much as the average result that we would have with a blind guess), but it is more relevant to focus on the ability to predict the trustworthiness of tags within some range, rather than the exact trust value. Depending on the choice of the threshold, the accuracy in this case varies in the range of 40% - 90%, as we can see in Figure 5.3. For thresholds higher than 0.85 (the most likely choices), the accuracy is at least 70%. We also compared the provenance-based estimates with the reputation-based ones, with a 95% confidence level Wilcoxon signed-rank test that proved that the estimates of the two algorithms is not statistically different. *For the* Waisda? *dataset, reputation- and provenance-based estimates are equivalent: when reputation is not available or it is not possible to compute it, we can substitute it with provenance-based estimates.* This is particularly important, as the availability of PROV data grows, one can compute trust values for data which is not associated with a trust value.

The "regression-by-discretization" approach consists in first a discretization of the continuous features at our disposal (e.g., timestamps) and a subsequent computation of regression by means of a classification algorithm (e.g., Support Vector Machines). If we apply it for making provenance-based assessments, then we approximate our trust values. This is not necessary with the reputation approach. Had we applied the same approximation to the reputations as well, then provenance-based trust would have performed better, as proven with a 95% confidence level Wilcoxon signed-ranked test, because reputation can rely only on evidence regarding the user, while provenance-based models can rely on larger data sets. Anyway, we have no need to discretize the reputation and, in general, we prefer it because of its lightweight computational overhead.

5.7 Combining reputation and provenance-based trust

Lastly, we provide a procedure for combining reputation- and provenance-based estimates to improve our predictions. If a certain user has been reliable so far, we can reasonably expect him/her to behave similarly in the near future. So we use reputation and we also constantly update it, to reduce the risk of relying on over-optimistic assumptions (if a user that showed to be reliable once, will maintain his/her status forever). However, reputation has an important limitation. To be reliable, a reputation has to be based on a large amount of evidence, which is not always possible. So, in case the reputation is uncertain, or in case the user is anonymous, other sources of information should be

used in order to correctly predict a trust value. The trust estimate based on provenance information, as described in Section 5.6, is based on behavioral patterns which have a high probability to be shared among several users. Hence, if a reputation is not reliable enough, we substitute it with the provenance-based prediction.

5.7.1 Procedure

The algorithm is as follows:

Algorithm 7: Procedure for combining reputation- and provenance-based trust.

```
1 Procedure provenance_reputation_estimation(artefact_provenance,artefact)
2 |  q_ev ← evaluate_user_evidence(user,artefact)
3 |  if q_ev > min_evidence then
4 |  |  trust_value ← predict_reputation(user,artefact)
5 |  else
6 |  |  trust_value ← predict_provenance(artefact_provenance, artefact)
7 |  end
8 |  return trust_value
```

evaluate_user_evidence This function quantifies the evidence. Some implementation examples: (1) *count*; (2) compute a subjective opinion and check if the uncertainty is low enough.

5.7.2 Application evaluation

We adopt the predictions obtained with each of the two previous procedures. The results are combined as follows: if the reputation is based on a minimum number of observations, then we use it, otherwise we substitute it with the prediction based on provenance. We run this procedure with different values for both the threshold and the minimum number of observations per reputation. We instantiate the *evaluate_user_evidence(user,artefact)* function as a *count* function of the evidence of *user* with respect to a given *tag*.

5.7.3 Results

The performance of this algorithm depends both on the choice of the threshold for the decision and on the number of pieces of evidence that make a reputation reliable, so we ran the algorithm with several combinations of these two parameters (Figure 5.3). The results converge immediately, after having set the minimum number of observations

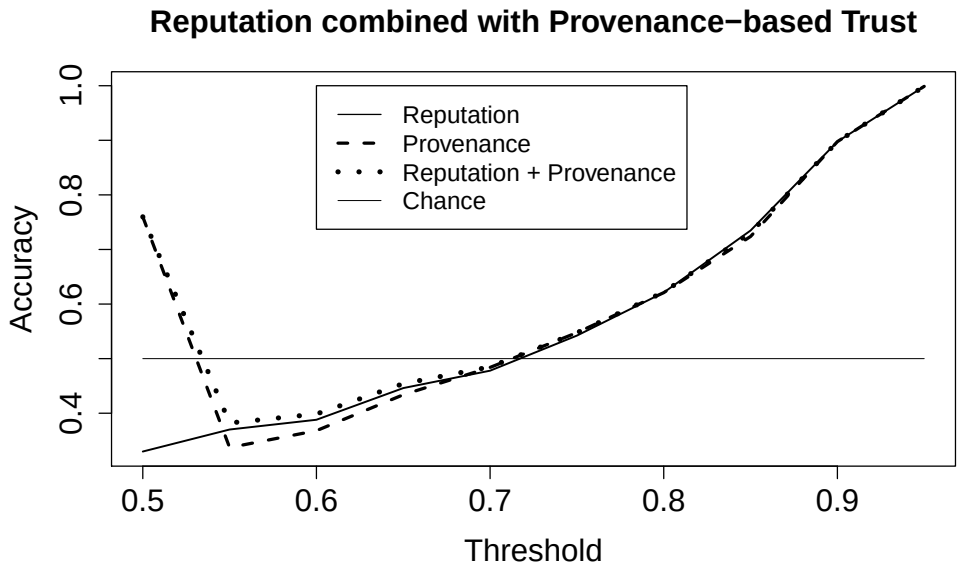

FIGURE 5.3: The figure shows accuracy varies for the different types of trust assessments.

at two. We compared these results with those obtained before. Two Wilcoxon signed-rank tests (at 90% and 95% confidence level with respect to respectively reputation- and provenance-based assessments) showed that *the procedure which combines reputation and provenance evaluations in this case performs better than each of them applied alone.* For thresholds higher than 0.85, the accuracy is at least 70%. Moreover, we would like to stress how the combination of the two procedures performs better than (in a few cases, equal to) each of them applied alone, regardless of the threshold chosen.

Combining the two procedures allows us to go beyond the limitation of reputation-based approaches. Substituting estimates based on unreliable reputations with provenance-based ones improves our results without significantly increasing risks, since we have previously proven that the two estimates are (on average) equivalent. Hence, when a user is new in a system (and so his/her history is limited) or anonymous, we can refer to the provenance-based estimate to determine the trustworthiness of his/her work, without running a higher risk of poor trust prediction. By performing a Pearson correlation test with a confidence level of 99%, we saw a small positive correlation between the reputation-based and provenance-based estimates. This implies that in some cases, reputation-based and provenance-based estimates behave alike. Thus we can substitute uncertain reputation-based assesments with the corresponding provenance-based assessments. This explains also the similarity among the results shown in Figure 5.3.

5.8 Conclusion

In this chapter, we first explored the correlation between user demographics and user reputations and showed the existence of such a correlation in the *Waisda?* dataset. Moreover, we showed how it is possible to use demographics extracted from user profiles to create user stereotypes (user abstractions based on demographics) and use them as a basis for trust estimation. However, in the *Waisda?* dataset user stereotypes were not useful to discriminate user reputation, although we found a correlation between single demographics (age, gender, etc.) and user reputation. Moreover, we showed how to use the FOAF ontology to represent both user profiles and stereotypes. Although the *Waisda?* dataset had a small number of registered users, we presented a methodology which will be extended and applied to `Steve.Museum` dataset in Chapter 7 and Chapter 8.

Additionally, we proposed and evaluated procedures for computing trust assessments based on reputation and based on provenance information, and for combining these two types of assessments. We showed that using reputation for trust assessment is simple, computationally light and accurate. We also showed the potential of provenance-based trust assessments: these can be at least as accurate as reputation-based methods and can be used to overcome the limitations of a reputation-based approaches (at least within a tagging environment). In the *Waisda?* dataset the combination of the two methods was more powerful than each of the two alone.

Chapter 6

Efficient Semi-automated Assessment of Annotations Trustworthiness

In this chapter we present efficient techniques to compute trust of annotations and annotators. The work in this chapter was presented at the 11th International Privacy, Security and Trust Conference (PST 2013) in Tarragona, Spain and won the "best student paper award ex-aequo". An extended version of this work was then published in a special issue of the Journal of Trust Management in 2014. My contributions are in the conceptualisation of Algorithm 9 along with the design and implementation of the clustering mechanism and co-designing Algorithm 8.

6.1 Introduction

The goal of this chapter is to develop algorithms for computing the trustworthiness of annotations in a fast and reliable manner. The techniques introduced in the previous chapter are able to compute the quality of annotations and reputation of annotators with good accuracy. However as the techniques are mainly evidence-based, the time taken for computation increases as the evidence increases. We introduce three new techniques for evaluating annotations for cultural artefacts. Firstly we present a technique which is a modification of techniques introduced in the earlier chapters for computing quality of annotations. It employs annotator trustworthiness to compute trustworthiness of their contributions, but instead of employing thresholds to accept or reject annotations, we devise a sorting mechanism. Accuracy of trust values is achieved by carefully handling the information at our disposal and by utilising the existence of a relationship between

the features considered (e.g., the annotation creator) and the trust values themselves. If the information is handled correctly and the relationship holds, then the trust values are accurate enough and form a basis to automatically decide whether or not to use the annotations. We evaluate this first algorithm by applying our algorithm on two different datasets, one from a `SEALINCMedia` project experiment and the other from the `Steve.Museum` dataset. In both cases we divide the dataset into two parts, training set and test set, so as to build a model based on subjective logic and semantic similarity in the training set, and then evaluate the accuracy of such a model on the test set.

Secondly we present a technique which makes it possible to perform trust estimations in a relatively fast manner, by properly clustering the training set on a semantic similarity basis. Here the goal of the contribution is to reduce the computational overhead due to avoidable comparisons between evaluated annotations and new annotations. We evaluate this contribution by applying clustering mechanisms in the training set data of the aforementioned datasets and by running our algorithm for computing trust values on the clustered training sets. The evaluation will check whether clustering reduces the computation time (and in case it does, up to which magnitude) and whether it affects the accuracy of the predictions.

Thirdly we present a technique for trust computations based on provenance information. In Chapter 5 we used information in annotators profile to group them into stereotypes. In this chapter we use other factors; for instance, certain periodic intervals, such as the time of the day or day of the week. Being able to recognise such stereotypes, we can compute a reputation per stereotype rather than per user. This approach guarantees the availability of evidence, as typically multiple users belong to the same stereotype and also compensates for the lack of evidence about specific users. We evaluate our hypothesis over the two datasets mentioned before by splitting them into two parts, one to build a provenance-based model and the other to evaluate it.

6.2 Related work

In Cilibrasi et al. [20], hierarchical clustering is used for grouping related topics, while Ushioda et al. [98] experiment on clustering words in a hierarchical manner. Begelman et al. [32] present an algorithm for the automated clustering of tags on the basis of tag co-occurrences in order to facilitate more effective retrieval. A similar approach is used by Hassan-Montero and Herrero-Solana [43]. They compute tag similarities using the Jaccard similarity coefficient and then cluster the tags hierarchically using the k-means algorithm. In our work, to build user reputations, we cluster the tags along with their respective evaluations (e.g., accept or reject). Each cluster is represented by a medoid

(that is, the element of the cluster which is the closest to its center), and in order to evaluate a newly entered tag by the same user, we consider clusters which are most semantically relevant to the new tag. This helps in selectively weighing only the relevant evidence about a user for evaluating a new tag.

6.3 Datasets adopted

We validate the algorithms we propose over two datasets of annotations of images[1]. The annotations contained in these datasets consist of content descriptions and the datasets contain also the evaluations from the institutions that collected them. For each annotation, the datasets contain information about its author and a timestamp. Since each institution adopts a different policy for evaluating annotations, we try to learn such a policy from a sample of annotations per dataset, and find a relationship between the identity of the author or other information about the annotation and its evaluation. We use the `Steve.Museum` dataset which we described in detail in Chapter 2 and for our experiments we considered only "usefulness-useful" as a positive evidence while all other categories are considered as negative evidence. The tags classified as "todo" are discarded. We also use the `SEALINCMedia` dataset which we describe in detail as follows.

`SEALINCMedia` dataset

The `SEALINCMedia` dataset was used for our experiments. This experiment used to obtain the dataset is described by Mieke et al. [64]. A total of 2,650 annotations resulted from the experiments, and these were manually evaluated by trusted personnel for their quality and relevance using the following scale:

- 1 : Irrelevant
- 2 : Incorrect
- 3 : Subjective
- 4 : Correct and possibly relevant
- 5 : Correct and highly relevant
- typo : Spelling mistake

[1]Complete results at `https://github.com/anottamkandath/Datasets/tree/master/Chapter6`

These tags, along with their evaluations, were used to validate our model. For each tag, the `SEALINCMedia` dataset presents the following elements: author identifier, artefact identifier, timestamp, evaluation. We do not focus on the goals of the experiment from which this dataset is obtained, that is, we do not analyse the relation between the kind of tag that was proposed to the user, and the tag that the user provided. We focus on the tag that the user actually proposes and its evaluation and we try to predict the evaluation of the tags provided by each user, given a small training set of sample evaluations about each of them.

We neglect the tags evaluated as "Typo" because our focus is on the semantic correctness of the tags, so we assume that such a category of mistakes would be properly avoided or treated (e.g. by using autocompletion and checking the presence of the tags in dictionaries) before the tags reach our evaluation framework. We build our training set using a fixed amount of evaluated annotations for each of the users, and form the test set using the remaining annotations. The number of annotations used to build the reputation and the percentage of the dataset covered is presented in Table 6.1: in the first column "# tags per reputation" we report the number of evaluated annotations we use to build each reputation, while in the second column, "% training set covered" we report the percentage of annotation used as training set compared to the whole dataset.

6.4 High-level system overview

The system that we propose aims at relieving the institution personnel (reviewers in particular) from the burden of controlling and evaluating all the annotations inserted by users. The system asks for some interaction with the reviewers, but tries to minimise it. Figure 6.1 shows a high-level view of the model.

For each user, the system asks the reviewers to review a fixed number of annotations, and on the basis of these reviews it builds user reputations. A reputation is meant to express a global measure of trustworthiness and accountability of the corresponding user. The reviews are also used to assess the trustworthiness of each tag inserted afterwards by a user: given a tag, the system evaluates it by looking at the evaluations already available. The evaluations of the tags semantically closer to the one that we evaluate have a higher impact. So we have two distinct phases: a first training step where we collect samples of manual reviews, and a second step where we make automatic assessments of tags trustworthiness (possibly after having clustered the evaluated tags, to improve the computation time). The more reviews there are, the more reliable the reputation is, but this number depends also on the workforce at the disposal of the institution. On the other hand, as we will see in the following section, this parameter does not affect

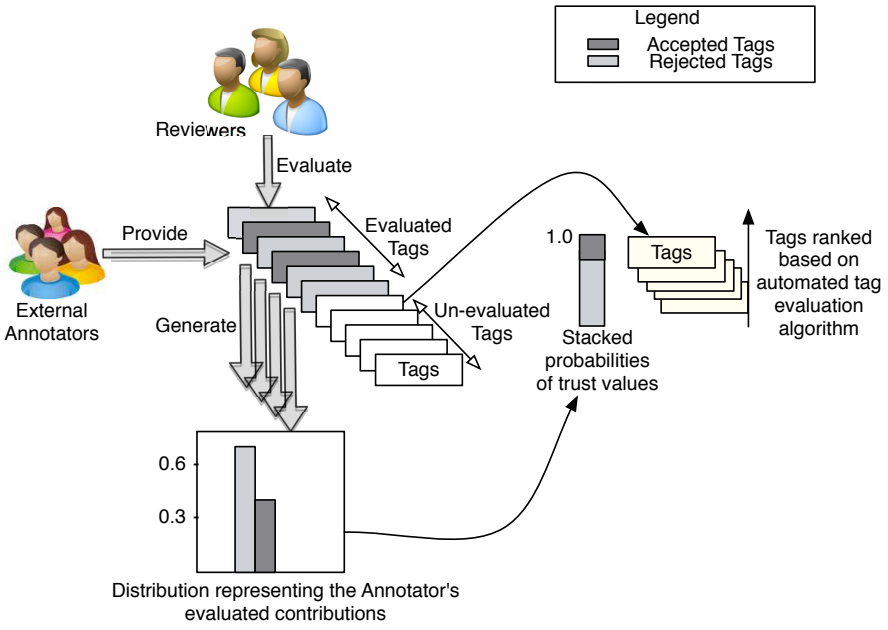

FIGURE 6.1: High-level overview

significantly the accuracy obtained. Moreover, we do not need to set an "acceptance threshold" (e.g. accept only annotations with a trust value of say at least 0.9, for trust values ranging from zero to one), in contrast to methodology in Chapter 3 and Chapter 5. This is important since such a threshold is arbitrary, and it is not trivial to find a balance between the risk to accept wrong annotations and to reject good ones.

6.5 Modified tag evaluation technique

The algorithm that we present here is similar to the ones introduced in Chapters 3, 4 and 5 for evaluating annotations. The novelty in this technique is that in order to evaluate tags (i.e. decide to accept or reject them), we define an ordering function on the set of tags based on their trust values (see Equation (6.1)). The ordered set of tags is represented as $\{t\}_1^{|tags|}$, where $|tags|$ is the cardinality of the set of tags. For tags t_1 and t_2,

$$t_1 \leq t_2 \iff E(\omega_{t_1}^m) \leq E(\omega_{t_2}^m) \tag{6.1}$$

From Chapter 2, we know that $E(\omega_u^m)$ is the user reputation, being the expected percentage of correct tags created by the user. Hence, we accept the last $E(\omega_u^m) \cdot |tags|$ tags in $\{t\}_1^{|tags|}$ (see Equation (6.2)) as $\{t\}_1^{|tags|}$ is in ascending order, so we accept the tags having higher trust value.

$$evaluation(tag) = \begin{cases} rejected & \text{if } t \in \{t\}^{1}_{E(\omega_u^m)\cdot|tags|} \\ accepted & \text{otherwise} \end{cases} \tag{6.2}$$

We provide here a pseudocode representation of the algorithm that implements the tag evaluation procedures, and we explain it in detail.

Algorithm 8: Algorithm to compute trust values of tags base on user reputation.

Input: A finite set of elements in $Training\_set = \{\langle tag, evaluation, UserID\rangle\}$ and $Test\_set = \{\langle tag, UserID\rangle\}$

Output: A finite set of evaluated tags $Result\_Test\_set = \{\langle tag, trust\_values\rangle\}$

```
1  for UserID ← UserID_1 to UserID_n do
2  |                                  ▷ for all tags in Training_set
3  |   rep[UserID] ← build_reputation(Training_set)
4  for UserID ← UserID_1 to UserID_n do
5  |                                  ▷ for all users in Test_set
6  |   for Tag ← tag_1 to tag_n do
7  |   |                              ▷ for all tags in Test_set
8  |   |   trust_values[Tag] = comp_tv(Training_set)
9  |   s_tags ← sort(tags(trust_values)
10 |   Result ← assess(s_tags,rep[UserID])
11 return Result
```

Input

The algorithm takes as input two vectors. The first vector, i.e. the training set, is composed of tuples formed by tags, their evaluation (e.g. "useful") and the user identifier (which consists of a URI, since we use the Semantic Web representation described above). The second vector (test set) is composed of tuples formed by tags and the identifier of the user that provided them. The computations used in the algorithm are described in detail in Chapter 2.

Output

The intended output consists of a vector of tuples formed by the tags in the test set and their estimated evaluation.

build_user_reputation

Builds a reputation for each user in the training set, following Equation (2.7). A reputation is represented as a vector of probabilities for possible tag evaluations.

trust_values

Trust values are represented as vectors of probabilities of possible tag evaluations, following Equation (2.11).

comp_tv

Implements Equation (2.11) using Equation (2.12). The value actually stored is the expected value of the opinion, that is $E(\omega_t^m) = \frac{p_t^m}{p_t^m+n_t^m+2} + \frac{1}{2} \cdot \frac{2}{p_t^m+n_t^m+2}$.

sort_tags

The tags are sorted according to their trust value, following the ordering function in Equation (6.1).

assess

The assess function assigns an evaluation to the tag, by implementing Equation (6.2).

6.6 Clustering semantically related tags

Reputations built using large training sets are likely to be more accurate than those built using smaller ones. On the other hand, the larger the set of tags used for building the reputation, the higher the number of comparisons we will have to make to evaluate a new tag. In order to reduce this tension, we cluster the tags in the training set of a user based on semantic similarity, for each resulting cluster we compute the medoid (that is, the element of the cluster which is, on average, the closest to the other elements), and we record the evidence counts. Clustering is performed on a semantic basis, that is, tags are clustered in order to create subsets of tags having similar meanings. After having clustered the tags, we adapt the algorithm so that we compute a subjective opinion per cluster, but we weigh it only on the semantic distance between the new tag and the cluster medoid. In this way we reduce the number of comparisons (we do not measure the distance between the new tag and each element of the cluster), but we still account for the size of the training set, as we record the evidence counts of it. We use hierarchical clustering (as described in detail in Chapter 2) for semantically clustering the words, although it is computationally expensive, because: (1) we know only the relative distances between words, and not their position in a simplex (the semantic distance

is computed as $1 - similarity(word_1, word_2)$), and this is one of the algorithms that requires such kind of input; and (2) it requires only one input argument, a real number "cut", that determines the number of clusters of the input set S of words. If cut = 0, then there is only one cluster; if cut = 1, then there are n clusters, where n is the cardinality of S. Clustering is performed offline, before any tag is evaluated, and here we focus on the improvement of the performance of the newly introduced tags. Algorithm 9 incorporates these optimizations. As Algorithm 8, Algorithm 9 takes as input the training set (composed of tuples formed by a tag, its evaluation and its author identifier) and a test set (composed of tuples formed by tags and their author identifier) and outputs a set of tuples formed by the tags in the test set and their estimated evaluations.

Algorithm 9: Algorithm to compute trust values of tags based on user reputation, with clustering of the evaluated tags in the training set.

Input: A finite set of elements in $Training\_set = \{\langle tag, evaluation, UserID\rangle\}$ and $Test\_set = \{\langle tag, UserID\rangle\}$
Output: A finite set of evaluated tags $Result\_Test\_set = \{\langle tag, trust\_values\rangle\}$

```
1  for UserID ← UserID_1 to UserID_n do
2  |                                  ▷ for all tags in Training_set
3  |   rep[UserID] ← build_reputation(training_set)
4  |   clusters[UserID] ← build_clust(training_set)
5  |   medoids[UserID] ← get_med(clusters,UserID)
6  for UserID ← UserID_1 to UserID_n do
7  |                                  ▷ for all users in Test_set
8  |   for Tag ← tag_1 to tag_n do
9  |   |                                  ▷ for all tags in Test_set
10 |   |   trust_values[Tag] = comp_tv(medoids[UserID],rep[UserID])
11 |   sort_tags ← sort(trust_values)
12 |   Result ← assess(sort_tags,rep[UserID])
13 return Result
```

6.7 Provenance-based trust values

The algorithms described so far are based on the fact that there exists a relationship between the identity of an author and the trustworthiness of his annotations, or that the user reputation is a meaningful estimate. However, there might be cases when the user reputation is not available, for instance if there is not enough evidence about his trustworthiness or in case his identity is not known. We show that the algorithm is not firmly dependent on the user reputation and, in case this is not available, other classes of information can be used as well. This class of information is so-called provenance information about how an artefact (in this case, an annotation) has been produced,

and represents, therefore, an extension of the information about the sole author of the annotation.

We follow a reasoning similar to methodology introduced in Chapter 5 as we use "provenance stereotypes" to group annotations. By stereotype we mean a class of provenance traces classified according to the user behaviour they hint at. For instance, we could have "Monday early morning users" or "Saturday night users". We suppose that a given behaviour should be associated with a particular reputation and hence with a given degree of trustworthiness of the annotations created in that manner, for two reasons:

- The trustworthiness of a given annotation might be affected by when it is created. For instance, late at night, users may on average be more tired and hence less precise than on other moments of the day.
- Users tend to follow a regular pattern in their behaviour, because, for instance, their availability for annotating is constrained by their working time. Therefore, by considering their behaviour, we implicitly consider their identity as well, even when they act as anonymous users.

In order to apply this kind of reasoning, we need to refer to the provenance information at our disposal about the annotations. In particular, these include only the day of the week and the time of creation for the dataset considered, but other information, when available, might be used as well (e.g. the typing duration for a given annotation). Since annotations are hardly created at the same time, in general do not coincide, we need to group them in order to be able to identify patterns in the data that allow us to link specific provenance information to the trustworthiness of the tags. In fact, the creation time of a tag may be recorded as a timestamp, but since tags are probably created at different times, we need to increase the granularity of this piece of information and analyse the part of the day or the day of the week when the tag was created, rather than the exact moment (tracked by the timestamp). In the datasets used in this chapter, the timestamps are the server times given in absolute time. Of course, this grouping introduces some uncertainty in the calculations because it introduces an approximation and because, in principle there are several possible groupings that we can apply, with different granularity and semantics (e.g. the days can be distinguished in weekdays and weekends, or simply be kept as single days of the week). In the next section, we report the results we obtained and we provide a possible explanation of why the grouping we propose allowed us to obtain the results we achieved, in the case studies we analysed. Lastly, from the modelling point of view, each group or stereotype can be thought of as a **PROV:BUNDLE** from the PROV Ontology as described in Chapter 2, that is a "named

set of provenance descriptions", where each set groups provenance traces according to the day of the week and the part of the day they belong to.

Unlike the methodology in Chapter 5, we do not apply support vector machines to learn the trustworthiness of the annotations created with a given stereotype. Rather, we collect a predefined amount of evidence (i.e. of evaluated annotations) per group, and we evaluate the remaining annotations of the same group based on the reputation estimated using the evidence collected, so as to exploit the provenance semantics instead of using it only as a statistical feature.

For representing provenance information we adopt the W3C Recommendation PROV-O Ontology described in Chapter 2.

Computing the reputation of a provenance stereotype

Once we have decided how to group the provenance traces, we start collecting evidence per group. We fix a limit to the amount of evidence needed to create the opinion representing the stereotype's reputation. (In the experiment described in the next section we vary this limit to evaluate the impact it has on the accuracy of the reputation itself.) The reputation is computed as in the **build_reputation()** procedure described in Algorithm 10. First we determine which stereotype the annotation belongs to. Then we increment the evidence count for the evaluation of the current tag until we reach the limit per stereotype. Lastly, we convert the list of evidence counts in subjective opinions.

Once the training set has been built, we evaluate the trustworthiness of the annotations in the test set for each group. We compare each annotation to be evaluated against each piece of evidence in the training set, and we use the semantic similarity emerging from that comparison to weigh the evidence and compute an opinion per annotation.

Once we have obtained one trust value per tag, we have to decide whether or not to accept the tag itself. To be more precise, for each tag we compute an entire opinion, representing the probabilities for each tag to be correctly evaluated with one of the possible evaluations. Now we must decide which evaluation to assign to the annotation. One strategy would use, for each annotation, the evaluation having the higher probability. We do not adopt this strategy because by doing so we will most likely tend to evaluate all tags of a given stereotype with the same dominant evaluation. For instance, if 95% of the training set annotations of one stereotype are useful, we will most likely evaluate all its annotations in the test set as useful. In turn, this implies that we do not take into account that we estimated that 5% of the annotations are not useful.

So we use an approach that combines the stereotype reputation with the trust values of the annotations, because we want to take fully into account the probabilities that are estimated by means of the reputation, and trust values estimate the trustworthiness of annotations.

Algorithm 10 presents the algorithm for annotation evaluation. First, it provides a procedure for computing the reputation of provenance stereotypes that takes as input a training set composed of tuples formed by tags, their evaluation and the identifier of the provenance stereotype they belong to. This procedure returns a set of pairs consisting of provenance stereotype identifiers and their reputation. Then the algorithm evaluates the new annotations, i.e. the annotations in the test set. This second procedure takes as input the training set (formed by tuples composed of tags, their evaluations and the identifier of the provenance stereotype they belong to) and the test set (formed by tags and their provenance stereotype identifier) and outputs a series of pairs consisting of the list of tags in the test set and the corresponding predicted evaluations.

The functions used in Algorithm 10 are as follows:

Input

The algorithm takes as input two vectors. The first vector, i.e. the training set, is composed of tuples formed by tags, their evaluation and provenance identifier (which indicates the stereotype group the tag belong to). The second vector (test set) is composed of tuples formed by tags and the identifier of the stereotype group they belong to.

Output

The intended output consists of a vector of tuples formed by the tags in the test set and their estimated evaluation.

build_reputation

The procedure to compute reputation per stereotype.

compute_reputation

Builds a reputation for each stereotype group in the training set, following Equation (2.7). A reputation is represented as a vector of probabilities for possible tag evaluations.

trust_values

Trust values are represented as vectors of probabilities of possible tag evaluations, following Equation (2.11).

comp_tv

Implements Equation (2.11) using Equation (2.12). The value actually stored is the expected value of the opinion, that is $E(\omega_t^m) = \frac{p_t^m}{p_t^m+n_t^m+2} + \frac{1}{2} \cdot \frac{2}{p_t^m+n_t^m+2}$.

sort_tags

The tags are sorted according to their trust value, following the ordering function in Equation (6.1).

assess

The assess function assigns an evaluation to the tag, by implementing Equation (6.2).

6.8 Implementation

The code for the representation and assessment of the annotations with the Open Annotation model has been developed using the SWI-Prolog Semantic Web Library[2](for `SEALINCMedia` dataset) and the Python libraries rdflib[3](for `Steve.Museum` dataset) and hcluster[4].

6.9 Results and discussion

We evaluated the algorithms that we proposed by running them on `Steve.Museum` dataset and `SEALINCMedia` experiment datasets. As described before, we split each dataset into a training and a test set, learn a model based on the training set, and evaluate it on the test set. There is a tradeoff between complexity and performance. On the one hand, a larger training set in general produces a more accurate model. On the other hand, an increased size of the training set induces a larger number of comparisons for each estimate, and

[2] `http://www.swi-prolog.org/pldoc/package/semweb.html`
[3] `http://www.rdflib.net/`
[4] `http://scipy-cluster.googlecode.com/`

Algorithm 10: Algorithm to compute trust values of tags using provenance stereotypes. First we present the procedure for computing the reputation of the provenance stereotypes and then we predict the trustworthiness of tags based on their provenance group.

```
1  procedure build_reputation()
       Input: A finite set of elements in
              Training_set = {⟨tag, evaluation, ProvenanceID⟩}
       Output: A set of provenance group reputations
              Result_Test_set = {⟨ProvenanceID, reputation_values⟩}
2      for tag in training_set_tags do
3          i ← tag.get_stereotype_id()
4          if length(trainingset[stereotypes[i]])<n then
5              trainingset[length(trainingset[stereotypes[i]]) + 1] ← get_eval(tag)
6          else
7              testset[length(testset[stereotypes[i]]) + 1] ← get_eval(tag)
8      for s in stereotypes do
9          rep[s] ← compute_reputations(s)
10     return s
   Input: A finite set of elements in Training_set = {⟨tag, evaluation, ProvenanceID⟩}
          and Test_set = {⟨tag, ProvenanceID⟩}
   Output: A finite set of evaluated tags Result_Test_set = {⟨tag, trust_values⟩}
1  for s in trainingset[stereotypes] do
2      rep[s] ← build_reputation(Training_set)
3  for s in testset[stereotypes] do
4      for Tag ← tag_1 to tag_n do
5          trust_values[Tag] ← compute_tv(Training_set)
6      s_tags ← sort_tags(trust_values)
7      Result ← assess(s_tags,rep[s])
8  return Result
```

hence an increased computation cost. To determine an optimal size for the training set in each case study, we have run the algorithm with different training set sizes, expressed in terms of annotations per user reputation, and tracked their performance.

Some errors can be due to intrinsic limitations of the experiment rather than imprecision of the algorithms. For instance, since training and test set are part of the same dataset, a larger training set means a smaller test set, and vice versa. Since our prediction is probabilistic, a small training set forces us to discretize our predictions, and this increases our error rate. Also, while an increase of the number of annotations used for building a reputation produces an increase of the reliability of the reputation itself, such an increase has the downside to reduce our test set size, since often only few annotators produce a large number of annotations. Nonetheless, we are bound to this limitation because we can only rely on learning reputations and trust values from museum evaluations since we

do not have any possibility to decide if the internal inconsistency of the tags regarding a given image implies low trustworthiness of one or more of them.

Both the `Steve.Museum` dataset and the `SEALINCMedia` dataset present an unbalanced distribution of tags. In each of the datasets, about three quarters of the tags are evaluated in a positive manner. The algorithm is developed in such a manner that, even if an annotator has a very high reputation (e.g. 95%), still we do not accept all his tags, rather we accept only the 95% of them. New tags are all classified as trustworthy only if the user reputation is 100% or if it is very high (e.g. 99%) and because of discretization, the amount of untrustworthy tags is so small (e.g. 2%) that it is neglected. So, it may happen that all the tags provided by a given user are predicted to be trustworthy, but since users are treated as "silos", i.e. they are evaluated independently of each other in our system, then this means that there are other users in the dataset for which some tags are predicted to be untrustworthy, so to justify an overall percentage of trustworthy tags.

Another important fact is that we cannot evaluate our system on a test set that is artificially balanced in terms of amount of positive and negative evidence. If we build the test set so that it is balanced, then our system will not be able to properly classify all the tags. Instead, we prefer to work with real data, so to be able to test if the annotator reputation is really representative of his performance. Since all the users in our system have high reputation, then necessarily our test set is unbalanced. Lastly, we must add that, since our system hardly evaluates all the tags as trustworthy, if the system was not able to predict at least some of the real trustworthy tags as trustworthy and some untrustworthy tags as untrustworthy, then the accuracy of the system would be higher than the percentage of "useful" tags from both datasets. The fact that this is not the case, as we will see in the remainder of this section, testifies the effectiveness of the algorithms proposed.

6.9.1 Estimation of annotation trustworthiness based on user reputation - Algorithm 8

First, we evaluated the performance of algorithm 8. The results of `SEALINCMedia` experiment are reported in Table 6.1, where correct tags are considered as a target to be retrieved, so that we can compute accuracy, precision, recall and F-measure. This first case study provided us interesting insights about the model that we propose. The evaluation shows positive results, with an accuracy higher than 80% and a recall higher than 85%.

Then, we applied the same evaluation over the `Steve.Museum` dataset and we reported the results obtained in Table 6.2, using the same metrics as before (that is, precision, recall, accuracy and F-measure). Here the performance is less favourable than for the first case study (accuracy around 70% and precision around 80%). This is possibly due to the different size of the `Steve.Museum` dataset, which may make it more varied than the `SEALINCMedia` dataset. Moreover, the basic assumption of our algorithm is the existence of a correlation between the user identity and his trustworthiness. This might not always be the case, or the correlation might not have always the same strength (e.g. a good user in some situations might not annotate accurately). Also, we aim at learning the museum standards for trusting annotations, but these are not always easy to learn. Lastly, the decrease of accuracy with respect to the previous case is possibly due to the different tag distribution (of positives and negatives) of the dataset and different domains. Different distributions can make it harder to discriminate between trustworthy and untrustworthy tags (as one may encounter mostly one type of observations). Different domains can lead to a different variability of the topics of the tags and this fact affects the reliability of clusters computed on a semantic basis (since clusters will tend to contain less uniform tags, and medoids will be, on average, less representative of their corresponding clusters), and consequently affects the accuracy of the algorithm. For e.g., both in the `Steve.Museum` and the `SEALINCMedia` datasets, the annotators were recruited through publicity in news papers and social media. Due to this, a large population provided more general annotations for describing the annotations. If there are targeted campaigns or calls for annotators with expertise in particular topics, then more diverse and detailed annotations would be provided by them.

It is important to stress that, on the one hand, the increase of the size of the training set brings an improvement of the performance, while on the other hand, performance is already satisfactory with a small training set (five observations per user). Also, this improvement is small. This is important because: (1) the sole parameter that we did not set (i.e. size of the training set) does not seriously affect our results; and (2) when the size of the training set is small, the performance is relatively high, so the need of manual evaluation is reduced. The results are satisfactory even with a small training set, also thanks to the smoothing factor of subjective logic that allows us to compensate for the possibly limited representativity (with respect to the population) of a distribution estimated from a small sample.

Results of the evaluation of Algorithm 8 over the `SEALINCMedia` dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

TABLE 6.1: Results of the evaluation of Algorithm 8 over the `SEALINCMedia` dataset.

# tags per reputation	% training set covered	accuracy	precision	recall	F-measure	time (sec.)
5	8%	0.73	0.88	0.81	0.84	87
10	19%	0.76	0.87	0.84	0.86	139
15	31%	0.76	0.86	0.86	0.86	221
20	41%	0.84	0.87	0.96	0.86	225

TABLE 6.2: Results of the evaluation of Algorithm 8 over the `Steve.Museum` dataset.

# tags per reputation	% training set covered	accuracy	precision	recall	F-measure	time (sec.)
5	18%	0.68	0.79	0.80	0.80	1254
10	27%	0.70	0.79	0.83	0.81	1957
15	33%	0.71	0.80	0.84	0.82	2659
20	39%	0.70	0.79	0.84	0.81	2986
25	43%	0.71	0.79	0.85	0.82	3350
30	47%	0.72	0.81	0.85	0.83	7598

Results of the evaluation of Algorithm 8 over the `Steve.Museum` dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

6.9.2 Improving computational efficiency of the estimation of annotation trustworthiness - Algorithm 9

We evaluated the performance of Algorithm 9 on both datasets. Table 6.3 and Table 6.4 report the results for the `SEALINCMedia` and the `Steve.Museum` datasets, respectively. Algorithm 9 is a variant of Algorithm 8 as it attempts to improve the computational efficiency of the first, while trying not to compromise its performance. We ran our evaluation with the same setting as before, with the same training set sizes. Moreover, in one case (Table 6.3) we also ran the algorithm with two different values for the "cut" parameter, to check its influence on the overall performance.

By comparing Table 6.3 with Table 6.1 we can see how the performance of Algorithm 8 is kept, and in some cases even improved, while the execution time is significantly reduced. The same holds for the `Steve.Museum` case, as we can see by comparing Table 6.4 and Table 6.2. Here, in a few limited cases the performance degrades, but in a negligible manner. The "cut" parameter, apparently, does not affect the performance much.

These considerations make us conclude that, at least in these case studies, it is worth clustering the training set on a semantic similarity basis, as this leads to a better computational efficiency, without compromising the performance in terms of precision, accuracy and recall.

TABLE 6.3: Results of the evaluation of Algorithm 9 over the `SEALINCMedia` dataset.

# tags per reputation	% training set covered	accuracy	precision	recall	F-measure	time (sec.)
clustered results (cut=0.6)						
5	8%	0.73	0.88	0.81	0.84	43
10	19%	0.82	0.87	0.93	0.90	24
15	31%	0.83	0.87	0.95	0.91	14
20	41%	0.84	0.87	0.96	0.91	18
clustered results (cut=0.3)						
5	8%	0.78	0.88	0.88	0.88	43
10	19%	0.82	0.87	0.93	0.90	14
15	31%	0.84	0.87	0.95	0.91	16
20	41%	0.84	0.87	0.96	0.92	21

Results of the evaluation of Algorithm 9 over the `SEALINCMedia` dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

TABLE 6.4: Results of the evaluation of Algorithm 9 over the `Steve.Museum` dataset.

# tags per reputation	% training set covered	accuracy	precision	recall	F-measure	time (sec.)
clustered results (cut=0.3)						
5	18%	0.71	0.80	0.84	0.82	707
10	27%	0.70	0.79	0.83	0.81	1004
15	33%	0.70	0.79	0.84	0.82	1197
20	39%	0.70	0.79	0.84	0.82	1286
25	43%	0.71	0.79	0.85	0.82	3080
30	47%	0.72	0.79	0.86	0.82	3660

Results of the evaluation of Algorithm 9 over the `Steve.Museum` dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

6.9.3 Estimation of annotation trustworthiness based on provenance tereotypes - Algorithm 10

We evaluated the performance of Algorithm 10 on both datasets. Table 6.5 and Table 6.6 present the results for the `SEALINCMedia` and the `Steve.Museum` datasets. We ran this evaluation with the same setting as before. Since we were interested only in checking whether the trustworthiness estimations based on provenance stereotypes perform as well as those based on user reputations in terms of precision and recall, we do not report the execution time of the algorithm.

By looking at the results we see that the performance is very satisfactory. In Table 6.5 precision is about 88% and recall ranges between 73% and 88%. The decrease in accuracy for the training set built with 20 annotations per reputation is plausibly due to the fact that many provenance stereotypes do not have 20 or more annotations available, so these cluster cannot contribute to the overall accuracy measurement, while they did with 5, 10 or 15 annotations per reputation.

Moreover, the amount of evidence needed to make these assessments is low, as demonstrated by the percentage covered by the training set over the dataset. In Table 6.6 the performance is even higher than in Table 6.5. First, this is due to the existence of a correlation between the provenance group an annotation belongs to and its trustworthiness. Second, the fact that the provenance stereotypes that we considered for this experiment are 21(since there are 7 days in a week and each day is divided into three time slots from 00:00 to 09:00, 09:00 to 17:00 and 17:00 to 00:00), which is much less than the number of users, together with the unbalance between useful and non-useful annotations in the `Steve.Museum` dataset (the first are much more plentiful than the latter) compensates a collateral effect of smoothing. In fact, smoothing helps in allocating some probability to unseen events (for instance, possible future mistakes of good users). So, because of smoothing, we predicted the existence of non-useful annotations for users who actually did not produce them (the dataset contains only relatively few non-useful annotations). Since there are many more users than provenance stereotypes, this error is higher with user-based estimates, where there are many more smoothed probability distributions (one per author), which causes many more annotations to be wrongly evaluated as non-useful. On the other hand, with provenance stereotypes, this error was much more limited, because the corresponding smoothed reputations introduced fewer wrong non-useful evaluations. Still, we will continue employing smoothing, as these are posterior considerations based on the availability of privileged information about the test set (i.e. its evaluation), and smoothing allows to compensate the lack of this information. On the other hand, the specific `Steve.Museum` dataset possibly shows a limitation of smoothing.

TABLE 6.5: Results of the evaluation of Algorithm 10 over the `SEALINCMedia` dataset.

# tags per reputation	accuracy	% training set covered	precision	recall	F-measure
5	0.68	1.69%	0.88	0.73	0.80
10	0.71	3.35%	0.87	0.80	0.83
15	0.78	4.97%	0.88	0.88	0.88
20	0.72	6.45%	0.87	0.80	0.83

Results of the evaluation of Algorithm 10 over the `SEALINCMedia` dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

TABLE 6.6: Results of the evaluation of Algorithm 10 over the `Steve.Museum` dataset.

# tags per reputation	accuracy	% training set covered	precision	recall	F-measure
5	0.84	0.25%	0.84	0.99	0.90
10	0.84	0.45%	0.84	0.99	0.90
15	0.84	0.66%	0.84	0.99	0.90
20	0.84	0.86%	0.84	0.99	0.90
25	0.84	1.04 %	0.84	0.99	0.90
30	0.84	1.22 %	0.84	0.99	0.90

Results of the evaluation of Algorithm 8 over the `Steve.Museum` dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

In the previous section, we hypothesized that the time of creation of an annotation may implicitly affect its trustworthiness and that the users follow approximatively regular patterns in their behaviours. To support these statements, we made the following analyses:

- We computed the average of the user reputations per provenance group. The averages vary from 0.73 to 0.84 in the `Steve.Museum` dataset and from 0.75 to 0.91 in the `SEALINCMedia` case study. Each user that took part in the `SEALINCMedia` experiment, participated only once. Moreover, their contributions are concentrated in the mid part of the weekdays, so we could not make additional checks. In the `Steve.Museum` dataset, instead, we also run a series of Wilcoxon signed-ranked tests at 95% confidence level (since the data distribution is not always normally distributed, as shown by a Shapiro-Wilk test at 95% confidence level, we prefer not to use a student t-test), and we discovered that:

  - There is no significant difference within user reputations in the morning, afternoon, and night slots respectively across the week. For instance, we took the reputations in the morning slots for Monday, Tuesday, etc. and the Wilcoxon signed-rank test showed no significant difference. The same holds for the afternoon and the night ones.
  - There is a significant difference between the morning and the afternoon slots and the afternoon and night slots. Here we compared the series of reputations per slot across the week.
  - If we compare the averages of the reputations with respect to the days (for instance, considering the three slots of Monday versus the three slots of Tuesday, etc.) we see no significant difference.
  - There is no significant difference between weekends and weekdays.

  The first two points support our hypothesis because they show that actually there are some relevant differences between groups and actually these depend on the time of creation of an annotation. The third and the fourth point show that, at least in this case study, it is not useful to keep track of the day of the week when the annotation was created. On the other hand, the fact that we recorded the day of the week allowed us to check if there is any difference both among days and between weekend and weekdays, while if we started directly with this latter distinction, we could not have decreased the granularity.

- As we stated in the previous item, the average number of provenance groups a user contribution belongs to is 1 in the `SEALINCMedia` dataset. In the `Steve.Museum` dataset, instead, the average number of groups a user contributions belongs to is 1.17, variance 0.56. This means that most of the users' contributions belong to one group. So we can say that, approximatively, there exists a one-to-many relation that links the groups with the users: given a group, we can identify a group of users that provide annotations mostly in that group. This means that, when we analyse the annotations that belong to a given group, then we implicitly analyse the annotations produced by a group of users that annotate mostly in that time interval. So the provenance group acts as a proxy to this group of users, and hence, in practice, we analyse the annotations in that group based on the reputations of the users linked to that group. In principle, there may be a high variance among the users belonging to a given provenance group. However, in the case studies analysed in this chapter, this does not happen to be the case, since the variance of the users reputation belonging to a given group is low.

- In the `Steve.Museum` dataset, the variance of the user reputations ranges between 0.12 and 0.15. This shows that, even if the averages of user reputations per group

range between 0.73 and 0.84, the reputations are not sparsely distributed. Rather, within provenance groups users tend to be rather homogeneous in terms of reputation. The same holds for the `SEALINCMedia` case study, where the variance of user reputation per provenance group ranges between 0.004 and 0.01;

- The time that we used in our computation is the server time and the fact that, in principle, the annotations are collected worldwide, this might imply that our calculations are misleading. However, since: (1) as shown before, there is a consistent distinction between morning, afternoon and night reputations (which is determined by user performance, and users tend to contribute at fixed times), (2) the amount of tags annotated as "problematic-foreign" is very small (about 1.9%) and (3) the artefact annotated in the case study belong mainly to U.S. cultural heritage institutions, we assume that the annotations are approximatively provided by users in the same time zone or in the neighbouring ones.

When grouping the tags based on time, the choice between coarser and finer granularity is not trivial and, in general, affects the uncertainty of the final result. Grouping the tags at a coarser granularity allows to easily collect evidence for a given group and find a semantic justification for the differences between groups. If we find a difference between morning and afternoon tags, we can easily suppose (and possibly test) that this is due to the influence that different parts of the day have on the user conditions (tired, sleepy, etc.). If we find a difference between tags made at 8.00 a.m. and at 9.00 a.m., we may need additional information to justify semantically the reasons of such differences. On the other hand, a finer granularity may reveal to be useful to avoid to group together heterogeneous tags. All these are generic considerations, and the choice of the best granularity depends on the peculiarities of the single use case evaluated. In our cases, as is evident from the considerations above, we chose a coarser granularity for the hours of the day and a finer one for the days of the week, because this combination was the most significant and gave us the highest accuracy.

6.10 Conclusion

We presented an algorithm for automatically evaluating the trustworthiness of user-contributed annotations by using subjective logic and semantic similarity to learn a model from a limited set of annotations evaluated by an institution. Moreover, we introduce two extensions of this algorithm. The first extension makes use of semantic similarity to cluster the set of evaluated annotations at our disposal (training set) and hence improves the computational efficiency of the algorithm. The second extension

regards the possibility to adapt the algorithm to use provenance information instead of the user reputation as a basis for the trustworthiness estimations.

We evaluated each algorithm on two different datasets of annotations from the cultural heritage domain. The algorithm based on user reputation satisfactorily allows us to estimate the annotation trustworthiness with an accuracy of about 80% in one case and 70% in the other one. Clustering effectively helps in increasing the efficiency of the first extension, and the use of provenance information helps us to compute accurate estimates of annotations trustworthiness.

Chapter 7

Trust Predictions using Extended Feature Sets

In this chapter we use machine learning techniques to determine trustworthiness of annotator and quality of annotation based on their properties. The work in this chapter was presented at the 10th International Workshop on Uncertainty Reasoning for the Semantic Web at the 13th International Semantic Web Conference (ISWC 2014) in Riva del Garda, Italy. The full version is now under submission at the Journal of Data Semantics. My contributions are in the conceptualisation, designing methodology, implementation, experimentation and evaluation.

7.1 Introduction

In the Chapters 3, 4 and 6 we used actual evidence available from annotation systems to model the reputation of the annotators and to determine quality of annotations. In Chapter 5 we presented a technique to predict quality based on properties of annotators, some of which were directly available and some of which were derived. In this chapter we explore further the possibility to understand which kinds of properties are important in deciding trustworthiness of annotation and of annotator. Demographic properties of annotators are entered by them when they create a profile in the annotation system. We derive more properties about them based on their performance in the system. Annotation properties are derived both at the annotation level and also in comparison to the other annotations. Once the properties regarding the annotation and its annotator are obtained, we employ them for computing trustworthiness of both annotation and annotator using machine learning techniques, which use features to predict a target or a goal (which in our case is quality of annotation or reputation of annotator). Since many

features of the annotator and annotations readily available or can be easily extracted, we can use machine learning techniques for trust predictions. However, employing all features might not be useful. Therefore in this chapter we describe techniques to selectively filter relevant features to perform good predictions. We summarise the original contributions of this chapter as follows:

1. An annotation quality evaluation methodology that combines Semantic Web data description and best-practice enrichment with machine learning techniques; crowd annotations are syntactically and semantically normalized through transformation into Linked Data format and enrichment with external vocabularies and knowledge bases. Given an annotation assessment ground truth, properties of annotations and of annotators are used to train machine learning algorithms for the classification of annotations' quality.
2. A mechanism for predicting reputation of annotators based on their properties and employing machine learning techniques to perform prediction of annotator reputation.
3. We extend the methodology in Chapter 5 of employing annotator and annotation properties for quality predictions and apply it on the `Steve.Museum` dataset.

The study shows that our approach can effectively support the annotation evaluation process and contributes new insights to the problem of annotator reputation estimation.

7.2 Related work

In the previous chapters we employed semantic similarity measures of annotations for trust computations. In this chapter we employ more features of words that the annotation is comprised of such as presence of annotations in *WordNet*, Flickr, Wikipedia and DBpedia, which can be seen as an extended analysis from the previous method.

Enrichment of text using vocabularies has been done in different domains. In systems where the user is restricted to provide textual information with restricted number of characters such as Twitter, enrichment of the text was done to derive meaning from the text. Previous works studied the enrichment of these short pieces of content with information from external knowledge sources. In [1], a semantic enricher OpenCalais[1] was used to identify 39 different types of entities such as persons, events and products in the textual content in addition to linking the content to news articles. Using the

[1] `http://www.opencalais.com/`

identified entities, their linking strategies achieved an accuracy of 70-80%. In another work, enrichment of Twitter messages was done with information from the ACM Computer Classification System, a hierarchical knowledge base, to evaluate messages from computer scientists with respect to their DBLP publications [40]. They applied different spreading activation functions to incorporate relations between concepts but found no significant improvement compared to not using these relations. These works show that enrichment of text does increase the potential for subsequent analysis. Some other works which employed enrichment are [48] and [103]. In our work we also enrich content, i.e. annotations, using external knowledge bases. However, due to the nature of the (mostly single word) annotations, we also propose different factors for enrichment applicable to the cultural heritage domain such as enrichment using the Wikipedia page of the creator of a digital artefact.

In our paper, we identify relevant features of annotation and annotator to determine quality of annotation and reputation of annotators. Other fields where properties are identified for a performing a task is text classification .i.e., assigning predefined categories to free-text documents. Relevant features of a document, e.g. author or length, are used to automatically predict the correct category for that document. When expert judgements are available for a set of documents, a supervised document classification method can be used to train a model. Such a model is then applied to new documents to determine the category of those documents. The combination of machine learning with crowdsourcing has been studied extensively in [63]. Here we use Naives Bayes method for our predictions. Annotation properties have also been studied in the context of Wikipedia [104] and Twitter [95]. In order to obtain information about properties of annotations to be used for machine learning predictions, we perform enrichment of the annotations in Flickr, Wikipedia, DBPedia and *WordNet*. Annotation quality has been shown to be related to properties of the annotator. The impact of user information such as *age*, *gender*, *education* and other *demographics* in crowdsourcing tasks have been explored in [61]. They explored the relationship between worker characteristics and their work quality and showed a strong link between them. In this paper we continue in this direction and investigate the relationship between annotation quality and a more extensive set of user properties.

7.3 Methodology

This section details the building blocks of our methodology, aimed at supporting cultural heritage institutions with the assessment of a large number of crowdsourced artwork annotations. The methodology is based on Linked Data and machine learning techniques

and results in a model which can be used to semi-automatically assess annotation quality according to customized evaluation criteria.

Figure 7.1 shows the workflow of our methodology. The first step is transforming the dataset into Linked Data using Semantic Web technologies as described in Chapter 2. Once the transformation is performed, enrichment of data is done by identifying properties related to annotations and annotators from the dataset and employing information from external vocabularies and knowledge bases. We then determine feature sets which can be used for prediction of annotation quality and annotator reputation as described in Section 7.3.3. In Section 7.3.4, the next step of identifying feature sets used for machine learning training and predictions is discussed. After performing the predictions, we assess the performance of our methodology and present results of our analysis.

FIGURE 7.1: Overview of the methodology

7.3.1 Prerequisites

A dataset to be used for semi-automatic assessment following our methodology should contain relevant **annotation properties** and **annotator properties**. We provide a generalized description of relevant properties in Section 7.3.2. The dataset should contain both annotations for which the *quality* value and users for which the *reputation* is known in advance. The training and test set should be representative of the entire set.

7.3.2 Relevant annotation and annotator properties

We base our notion of relevance on a detailed qualitative study that we performed on the the `Steve.Museum` dataset.

Since annotations are in the form of text, the properties we identified apply to those of text.We found that the language of annotations is a critical property. All content in non-English languages, and all incorrectly spelled words, were reviewed as non-useful. Statements of a more subjective nature, such as *peaceful* and *elegant*, were also not regarded as useful. Based on the reviewers comments we also identified that annotations should represent terms other people would use to search; for example on the annotation *strong focal point* and *slop*, reviewers commented "not likely a term others would use to search". Overly generic terms such as *modern* and *work* were also seen as non-useful.

Based on these findings of this representative project we identified seven properties related to annotations that play a role in the evaluation process.

1. **language** whether the annotation is written in the right language(s)
2. **spelling**, whether the annotation is spelled correctly
3. **objectiveness**, whether the annotation is factual
4. **popularity**, whether other people would use such words to describe something
5. **vocabulary**, whether the annotation matches the desired (standard) vocabulary
6. **similarity**, whether such an annotation would be used to describe similar objects
7. **specificity**, whether the annotation is specific enough

To effectively assess the quality of annotations, we hypothesize that enrichment of annotations with the above listed properties will support the process of semi-automatically assessing the quality of annotations. The specific implementations of these properties can vary case-by-case and could use different information from external vocabularies and knowledge bases to suit their needs.

Annotator properties were not taken into account during the annotation evaluation process by reviewers for the `Steve.Museum` dataset and thus we cannot draw general conclusions about their effect on annotation quality. `Steve.Museum` experiment did however gather personal information that annotators could (optionally) fill in. These characteristics were deemed relevant with respect to the annotation quality. For our analysis we consider these characteristics, which are shown in Table 2.2 as relevant annotator properties. This can also be extended to other crowdsourcing projects where annotators provide information about themselves which can be used to determine the quality of annotations they provide.

7.3.3 Creating a minimal and optimal feature set

The enriched dataset, built based on chosen external vocabularies and knowledge bases, contains many properties. Not all properties are equally relevant for accurate predictions and some properties might even be redundant.

To identify relevant properties for prediction, appropriate pair-wise statistical tests need to be performed. These tests indicate whether there is any significant difference between the observed variable in different classes; e.g. whether differences in annotation quality

can be subscribed to the level of education. We then select only those properties with significant difference in statistical test results for building a minimal set of features for prediction.

Some of the properties might be redundant, for example age and year of birth. To resolve this redundancy of properties, a correlation-based feature subset selection algorithm is used, which evaluates the worth of a subset of properties by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred [41]. This results in an optimal feature set regarding the prediction variable.

7.3.4 Training the model

The classification of annotation quality is then performed by applying a machine learning algorithm, using the properties selected in the previous step, and the enriched dataset. We opted for Naives Bayes classifiers (discussed in Chapter 2) since they have been used extensively for text and document classification and are a simple, yet effective technique. In this paper we want to demonstrate the applicability of this methodology and therefore we used the default parameter settings for Naives Bayes classifiers and did not experiment with optimizing parameters.

The annotation dataset is split into a training set on which the model is trained, and a test set on which the model is evaluated. To overcome possible biases introduced by the splitting, standard approaches like n-fold cross validation is used: the dataset is randomly split in n subsets. The results from performing the validation n times is averaged where each time one subset is used for validation and the remaining subsets for training.

7.4 Reputation modelling and feature set descriptions

We introduced the `Steve.Museum` dataset in Chapter 2. For our experiments in this chapter, the review classes were grouped in four main categories: *useful*, *not useful*, *problematic* and *judgements*. Following Chapter 2 we transformed the resulting data in Linked Data format. In Section 7.4.1 we present our modelling approach for the reputation of users and in Section 7.4.2 we describe how we enriched the dataset.

7.4.1 Modelling annotator reputation in Steve.Museum

The reputation of annotators, a value not originally included in the `Steve.Museum` dataset, is calculated with an approach based on subjective logic for modelling the annotator reputation based on positive and negative evidence as discussed in detail in Chapter 2. We build reputation in order to determine if there is an effect of annotator properties on the annotator reputation in a simpler setting first before trying to understand the influence of annotator properties on annotation reputation for multiple categories.

Aggregating the annotations and their evaluations per annotator helps us estimate the reputation of the annotator in the system based on the total number of *useful* and *non-useful* annotations. The *non-useful* annotations consist of the annotations in categories which were not classified as useful. We computed reputation of an annotator based on *useful* and *non-useful* because it can be considered as positive and negative evidence.

7.4.2 Annotation and annotator features

Section 7.3.2 described seven relevant properties to enrich annotations. We now describe the concrete implementation of these properties using information from the following sources: Wikipedia, DBpedia, Flickr and *WordNet*. Multi-word annotations are space-delimited into individual words. We describe how we used these sources as heuristic for the properties listed in Section 7.3.2.

Language and spelling. Since the `Steve.Museum` experiment mainly comprises museums in the United States, the desired *language* of annotations is English. We try to check if words in the annotation are in English by searching for its occurrence in *WordNet* (for nouns and adjectives) or in DBpedia (for named entities). This is because *WordNet* contains words in English and this verification also covers the *spelling* property, since we check for an exact string match and erroneously spelled words will not occur in these matches.

Objectiveness. The *objectiveness* is modelled as the number of adjectives (which indicate subjectiveness) in the words of the annotation. This information is extracted by querying for each word in *WordNet*.

Popularity. Reviewers' comments in `Steve.Museum` illustrated the importance that annotations should be specified in terms other people would search for as a phrase. Therefore we approximate the *popularity* of an annotation by counting the number of images on Flickr that have been tagged with all words in the annotation in a bounded time period.

Vocabulary. Since detailed descriptions about the artworks were not available, we modelled the *vocabulary* property by verifying whether the occurrences of each individual word in the annotation are contained in the description of the artwork creator, which we extracted from DBpedia. The rationale here is that the description will likely describe (using the correct vocabulary) important works made by that creator. We preprocessed the creator string and used that to automatically query DBPedia. In case of a single match, that resource was chosen, and if the search resulted in more results, this was disambiguated manually. The disambiguation was done as follows. The corresponding Wikipedia pages of the results were read in detail for disambiguation. The fact that the creators had distinctive types of artworks and their year of birth and death were mentioned in `Steve.Museum` dataset, made it easier for the disambiguation using Wikipedia. This process is described in detail in Chapter 2.

Similarity and specificity. Similarity is a measure how many times an annotation would be used to describe similar objects and *specificity* measures the specificity of the annotation. The `Steve.Museum` dataset contains a collection of annotations on specific objects, namely artworks. As *similarity* measure we take the frequency of a word in the dataset as representative for the likeliness people use that word to describe artworks in general. The *specificity* is modeled as the maximum depth of any of the words in the annotation in the hierarchical *WordNet* tree.

TABLE 7.1: Annotation properties in the $\boldsymbol{F_a}$ feature set.

Features used for enrichment
words in annotation
words in *WordNet*
adjectives in annotation
matches in Flickr
matches in Wikipedia
matches in Wikipedia creator page
Maximum depth of words
Annotation day of week
Annotation hour of day
Frequency of annotation

From our analysis of registered and anonymous users we found a small difference between the two groups in terms of the length of the annotations and the day and time when the annotations was created. Therefore, we also considered as relevant the following properties: 1) the *number of annotation words*, under the assumption that the more specific they are, the more useful they will be; 2) the *weekday* (day of the week the annotation was created) and the *hour* (hour of the day the annotation were created), as they might indicate the focus and the time people can spend on creating the annotation, influencing the annotation quality.

7.5 Evaluation

This section describes the results obtained by the application of the methodology introduced in Section 7.3 on the enriched `Steve.Museum` dataset. Experiments are performed using prediction algorithms in WEKA[2] by performing training with different feature sets. We use the annotations in the dataset which had a single review or multiple same reviews for predicting annotation quality. The annotations evaluated as *judgement* were not included in the analysis, since the size of that class was too small to be used for training and predictions. The reputation of annotators were computed using methodology defined in Section 7.4.1. Section 7.5.1 describes the different feature sets used for the predictions. Section 7.5.2 describes how predictions were done about annotation quality and 7.5.3 describes how predictions of annotation reputation are performed.

7.5.1 Feature set selection

To further our understanding of the effect of annotation and annotator characteristics we experiment using different feature sets. To this end, we define four feature sets used for predicting **annotation quality** and two feature sets for predicting the **annotator reputation**.

7.5.1.1 Annotation quality feature sets

To assess the effect of annotation characteristics on the annotation quality we define feature set $\boldsymbol{F_a}$ as the set of all annotation properties stated in Table 7.1.

$\boldsymbol{F_a}$ = *properties in Table 7.1*

To assess the effect of characteristics of the annotator with respect to annotation characteristics we define feature set $\boldsymbol{F_{a+u}}$ as the set of all annotation properties, stated in Table 7.1, and all annotator properties, stated in Table 2.2.

$\boldsymbol{F_{a+u}}$ = *properties in Table 7.1 + Table 2.2*

Feature set $\boldsymbol{F_{a+u}}$ contains 22 features that could be redundant or non-relevant for our prediction purposes. Therefore we define a minimal feature set, $\boldsymbol{F_{min\_a}}$, containing only relevant properties, and an optimal set, $\boldsymbol{F_{opt\_a}}$, containing only relevant and non-redundant properties.

[2] `http://cs.waikato.ac.nz/ml/weka/`

The minimal feature set $\boldsymbol{F_{min\_a}}$ is defined as the set of all properties in $\boldsymbol{F_{a+u}}$ that are relevant with respect to the annotation quality (i.e. the review class (*useful*, *not useful* or *problematic*) provided by the reviewers).

In order to determine relevance of a property, we divide the dataset into three subsets of annotations, each corresponding to the evaluations obtained, and comparisons are made between those subsets. We consider a property relevant when it can be used to discriminate (to any extent) between two review classes. Therefore, for each property, we perform pairwise correlation tests using only the property values from two review classes. This results in three tests per property. We choose the correlation test metric based on the type of each property. To test categorical properties (e.g. *education*) we use the Pearson's Chi-Squared test and for interval properties (e.g. # *words in annotation*) the Wilcoxon rank sum test, both with a 99.5% confidence interval. $\boldsymbol{F_{min\_a}}$ contains all properties for which at least one of the three pairwise comparisons falls within the confidence interval. In the case of four properties all three comparisons resulted in values outside the confidence interval. To illustrate this, the p-values for "Museum visits" are 0.72, 0.66 and 0.57, where a value below 0.005 was required for inclusion.

$\boldsymbol{F_{min\_a}}$ = $\boldsymbol{F_{a+u}}$ - [Museum visits, # matches in Wikipedia creator page, gender, income]

The optimal feature set $\boldsymbol{F_{opt\_a}}$ for annotation quality prediction contains features selected by the feature subset selection algorithm in WEKA We use the CfsSubsetEval algorithm which evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. In our case, it favours subsets of features that are correlated with the review class and which have a low inter-property correlation. Applying the algorithm to the properties in $\boldsymbol{F_{a+u}}$ resulted in:

$\boldsymbol{F_{opt\_a}}$ = [Language, # words in *WordNet*, maximum depth of words, # matches in Flickr, frequency of annotation]

7.5.1.2 Annotator reputation feature sets

To assess the effect of characteristics of the annotator on the annotator reputation we define feature set $\boldsymbol{F_u}$ as the set of all annotator properties stated in Table 2.2.

$\boldsymbol{F_u}$ = *properties in Table 2.2*

Feature set $\boldsymbol{F_u}$ contains 12 properties that could be redundant or non-relevant for our purposes. The prediction variable, the annotator reputation, is however a continuous

variable with no clearly defined class boundaries. This removes the utility of the manual analysis and we only define the optimal feature set $\boldsymbol{F_{opt\_u}}$. Similar to the optimal feature set for annotation quality we define $\boldsymbol{F_{opt\_u}}$ as the output of the feature subset selection algorithm in WEKA (CfsSubsetEval). Applying the algorithm to the features in $\boldsymbol{F_u}$ resulted in:

$\boldsymbol{F_{opt\_u}}$ = [Education, internet usage, tagging experience]

7.5.2 Predicting annotation quality

We evaluate our approach for predicting the annotation quality using the Naive Bayes classifier in WEKA. We applied a 10-fold cross validation to the feature sets $\boldsymbol{F_a}$, $\boldsymbol{F_{a+u}}$, $\boldsymbol{F_{min\_a}}$ and $\boldsymbol{F_{opt\_a}}$. In this section we describe the classification results.

TABLE 7.2: Classification results for annotation quality prediction.

Feature set	Class	Precision	Recall	F-measure
$\boldsymbol{F_a}$	useful	0.913	0.896	0.904
	not-useful	0.175	0.193	0.183
	problematic	0.043	0.110	0.062
$\boldsymbol{F_{a+u}}$	useful	0.915	0.890	0.902
	not-useful	0.175	0.205	0.189
	problematic	0.113	**0.292**	0.163
$\boldsymbol{F_{min\_a}}$	useful	0.915	0.896	0.905
	not-useful	0.181	0.207	0.193
	problematic	0.126	0.263	0.170
$\boldsymbol{F_{opt\_a}}$	useful	0.913	0.913	**0.913**
	not-useful	0.199	0.206	**0.202**
	problematic	**0.397**	0.129	**0.195**

TABLE 7.3: Confusion matrix for annotation quality with the $\boldsymbol{F_a}$ feature set.

	useful	not-useful	problematic
useful	35840 (89.5%)	3818 (9.54%)	354 (0.88%)
not-useful	3258 (77.07%)	814 (19.25%)	155 (3.66%)
problematic	164 (78.46%)	22 (10.52%)	23 (11.00%)

TABLE 7.4: Confusion matrix for annotation quality with the $\boldsymbol{F_{a+u}}$ feature set.

	useful	not-useful	problematic
useful	35622 (89.02%)	4064 (10.15%)	326 (0.81%)
not-useful	3206 (75.80%)	868 (20.53%)	153 (3.61%)
problematic	119 (56.93%)	29 (13.80%)	61 (**29.18%**)

Table 7.2 shows the precision, recall and F-measure of predicting the categories of annotations for different feature sets. The confusion matrices for these feature sets $\boldsymbol{F_a}$,

TABLE 7.5: Confusion matrix for annotation quality with the $\boldsymbol{F_{min\_a}}$ feature set.

	useful	not-useful	problematic
useful	35838 (89.56%)	3914 (9.78%)	260 (0.64%)
not-useful	3230 (76.41%)	875 (**20.70%**)	122 (2.88%)
problematic	120 (57.41%)	34 (16.26%)	55 (26.31%)

TABLE 7.6: Confusion matrix for annotation quality with the $\boldsymbol{F_{opt\_a}}$ feature set.

	useful	not-useful	problematic
useful	36537 (**91.31%**)	3457 (8.63%)	18 (0.04%)
not-useful	3334 (78.87%)	870 (20.58%)	23 (0.54%)
problematic	130 (62.20%)	52 (24.80%)	27 (12.91%)

$\boldsymbol{F_{a+u}}$, $\boldsymbol{F_{min\_a}}$ and $\boldsymbol{F_{opt\_a}}$ are listed in Table 7.3, Table 7.4, Table 7.5 and Table 7.6 respectively.

From Table 7.2 we can observe that the highest F-measure for each class is achieved by using $\boldsymbol{F_{opt\_a}}$. There is a difference of less than 1%, 10% and around 68% compared to the $\boldsymbol{F_a}$ feature set for *useful*, *not-useful* and *problematic* classes respectively. The feature subset selection algorithm selected only five features for $\boldsymbol{F_{opt\_a}}$, whereas $\boldsymbol{F_a}$ has twice this number of features. From the results it seems that with more properties, the machine learning algorithm is less capable to generate a good prediction model. Precision and recall of the *useful* and *not-useful* classes remain stable over all feature sets, while the precision for the *problematic* class significantly increases for $\boldsymbol{F_{opt\_a}}$. We also observe that recall is highest (0.292) for the *problematic* class for the $\boldsymbol{F_{a+u}}$ feature set.

Tables 7.3, 7.4, 7.5 and 7.6 show the confusion matrices returned by the prediction algorithms, which shows the number of annotations predicted into different categories. For example, in Table 7.3, the first row indicates how the useful annotations were predicted. We see from the table that 35840 annotations were predicted correctly as useful while 3818 annotations and 354 annotations were wrongly predicted as not-useful and problematic respectively. Similarily the other rows indicate the predictions for the other categories. The largest amount of correct predictions for the *useful* class is achieved by using feature set $\boldsymbol{F_{opt\_a}}$ (91.3%), for *not-useful* by using feature set $\boldsymbol{F_{min\_a}}$ (20.7%) and for *problematic* by using feature set $\boldsymbol{F_{a+u}}$ (29.18%). The total number of annotations in the *useful* class is quite high compared to the number of annotations in the *problematic* class and due to this imbalance, the feature selection algorithms might tend to work better with annotations from *useful* class compared to the other classes.

The minimal feature set $\boldsymbol{F_{min\_a}}$ has 18 features compared to 22 features in the $\boldsymbol{F_{a+u}}$ feature set, which according our statistical tests were not relevant, and thus a similar performance was expected. Table 7.4 and Table 7.5 confirm this by showing a less than 3% difference for the *problematic* class. However, compared to these large features sets,

for the smaller feature set $\boldsymbol{F_{min\_a}}$ fewer instances were correctly classified as problematic (low recall). Also significantly fewer instances from other classes were classified as *problematic*. This indicates that there are features in the larger sets which the algorithm wrongly correlates with *problematic* class characteristics.

7.5.3 Predicting annotator reputation

The reputation of the annotators was not evaluated by `Steve.Museum`. We calculated reputation of registered annotators based on the review of their annotations using subjective logic as described in Chapter 2. The annotations which were evaluated as *useful* were considered as positive evidence. All other categories were considered as negative evidence.

Each registered annotator was assigned a *reputation* score, a value between 0 and 1, using the model described in Section 7.4.1. In this section we report the classification performance for users with **low** and **high** reputation. The bottom 100 registered users, having reputation scores between 0.09 and 0.68, represent the **low** class, and the top 100, with reputation scores between 0.88 and 0.98, represent the **high** class. We did not consider the other users for the analysis, since the goal of this analysis is to verify whether we can distinguish between these two classes.

Table 7.7 shows the classification performance and Table 7.8 shows the confusion matrix for both feature sets by performing a ten-fold cross validation. The confusion matrix for $\boldsymbol{F_{opt\_u}}$ is similar.The results are similar to a completely random approach, i.e. where the class is chosen randomly without using any other information. This shows that the *prediction power* of the user information gathered in the `Steve.Museum` project is very low and thus is not a good candidate for predicting the user reputation.

We also experimented by increasing the gap between low and high reputation users, by taking the first and last 75 and then 50 users. However, this did not result in a significantly better performance.

Table 7.7: Classification results for annotator reputation prediction.

Feature set	Class	Precision	Recall	F-measure
$\boldsymbol{F_u}$	High	0.51	0.56	0.54
	Low	0.51	0.46	0.48
$\boldsymbol{F_{opt\_u}}$	High	0.51	0.57	0.54
	Low	0.52	0.46	0.49

TABLE 7.8: Confusion matrix for annotator reputation with the $\boldsymbol{F_u}$ feature set. The confusion matrix for $\boldsymbol{F_{opt\_u}}$ is similar.

	useful	non-useful
useful	56(56%)	44(44%)
non-useful	54(54%)	46(46%)

7.6 Conclusion

We have presented a generic methodology to semi-automatically evaluate the crowdsourced annotations in cultural heritage domain. Our approach combines Semantic Web data descriptions and enrichment with machine learning techniques. We described conversion of crowdsourced data in cultural heritage to Linked Data, identification of relevant features which affect the quality of annotations and reputation of annotators and employing machine learning algorithms to perform predictions about annotation quality and annotator reputation.

Our methodology assumes availability of data regarding annotations, annotators and cultural heritage artefacts. The main goal of our approach is to identify and extract features both about annotation and about annotators which give an indication about their quality and to employ them for training and predictions.

The annotator properties were gathered from the `Steve.Museum` dataset. We tried to evaluate their relevance for determining quality of annotations. From our experiments, adding annotator properties along with annotation properties increases prediction results of *problematic* by 6.18% and of *not-useful* by 1.28%. Thus annotator properties help to increase prediction of smaller categories for the `Steve.Museum` dataset.

We also observed the effect of different feature sets on prediction accuracy. Using feature set $\boldsymbol{F_{opt\_a}}$ provides good results with an F-measure of 0.91 for *useful* annotations, while providing a low recall of 0.129 for the *problematic* class. This may be because the *useful* class comprises the majority of the dataset and thus the feature selection algorithm focuses on increasing the overall prediction results rather than the accuracy per category.

We studied in detail which features affect the annotation quality; the resulting minimal feature set $\boldsymbol{F_{min\_a}}$ was used for predictions. The results indicate that this feature set helps to predict the *useful* and *not-useful* categories better than the *problematic* class. We achieved highest F-measure of 0.91 for *useful* annotations using machine selected features. The factors which we considered relevant to predict annotation quality, such as language and spelling, objectiveness, popularity and similarity of annotations were indeed also considered relevant by the feature selection algorithm in WEKA as shown by the feature set $\boldsymbol{F_{opt\_a}}$. The annotation quality predictions were performed using

different feature sets. Also the properties of annotators together with properties of annotations help to predict smaller categories better, such as the *problematic* ones in the `Steve.Museum` dataset.

We predicted reputation of annotators based on their properties. The predictions of annotator reputation did not achieve better results than a random classification (of 50%). Our results indicate that we were not successful in predicting annotator reputation, since the performance was similar to that of a random prediction for high and low reputed annotators: the F-measure for *high* and *low* reputation annotators was approximately 0.5. In Chapter 2 we had shown that subjective logic is a good technique to model annotator reputation, and thus one of the reasons for this poor performance might be that the personal characteristics of an annotator as collected by `Steve.Museum` were simply not related to the quality of annotations he or she creates.

Chapter 8

Predicting Quality of Crowdsourced Annotations using Graph Kernels

In this chapter we employ graph kernels to make predictions about quality of crowdsourced annotations. The work in this chapter was presented at the IFIP Trust Management Conference 2015 in Hamburg, Germany. My contributions are in the conceptualisation, methodology, experimentation and evaluation.

8.1 Introduction

Properties of the annotations such as annotator, annotated artefact, time stamp etc. and properties of the artefact and of the annotators themselves can all be modelled using the Resource Description Framework (RDF) as described in detail in Chapter 2. Apart from representing the entities and the relations between them, such an RDF graph also captures the structural properties of the information.

As shown in Chapter 5 and Chapter 7, machine learning techniques such as Support Vector Machines can be used to make predictions about features in the dataset.

Recently, machine learning using graph kernels has arisen as an efficient method for learning from RDF graphs [26, 68], that can effectively exploit the structural properties of the graph using Support Vector Machines. To show the potential of such a graph kernel we apply it on the `Steve.Museum` dataset. First we transform the annotations and contextual information from the dataset to a semantic model and enrich the model with external vocabularies and knowledge sources as described in Chapter 2. We then leverage this model to make predictions about the annotation quality by applying the Weisfeiler-Lehman RDF graph kernel. The kernel computes the number of subtrees

shared between two graphs by using the Weisfeiler-Lehman test of graph isomorphism [92] and is explained in detail in Chapter 2.

In this chapter we utilize RDF graph kernels to utilize structural properties of graphs to make predictions about annotation quality. Although features about the user and of the annotations were used to make predictions of quality with Support Vector Machines in Chapter 7, we did not employ RDF graph kernels for the predictions. The work in this chapter aims to provide a new method employing RDF graph kernels for automatically predicting quality of crowdsourced annotations in the cultural heritage domain.

We show how a specialized kernel for RDF can be applied on a semantic cultural heritage annotation dataset to predict annotation quality and relevant features and provide insights into the benefit of RDF kernel for cultural heritage datasets.

8.2 Methodology

In this section we describe the workflow that we propose to assess the quality of crowdsourced annotations. We begin with an overview of the workflow and then we describe each component in detail.

8.2.1 Workflow overview

The workflow that we adopt to estimate the quality of the user-provided annotations is depicted in Figure 8.1 and consists of three steps:

1. Representing annotations in RDF
2. Annotations enrichment
3. Machine learning with graph kernels for RDF

Whenever an annotation is introduced in the system, it is modelled in RDF, along with its related metadata (e.g., its author). The resulting RDF graph is then enriched by linking it with information provided by authoritative and trusted Linked Data sources. In this manner, we expand the knowledge graph describing the annotation.

We transformed the `Steve.Museum` dataset into Linked Data using the model illustrated in Figure 2.2 and described in Chapter 2. Most properties of the users and the annotation could be mapped one-to-one. However, some annotations were reviewed multiple times. For the purpose of prediction we required each annotation to have exactly one review;

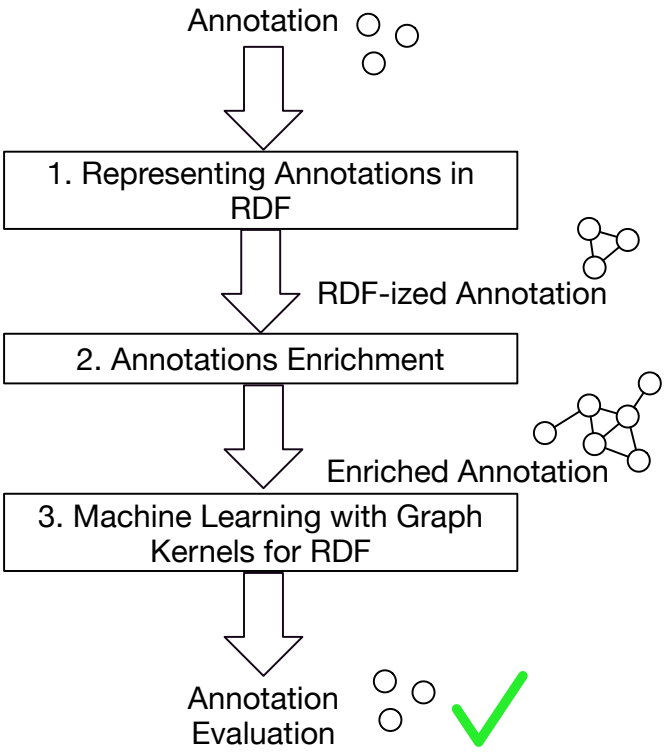

FIGURE 8.1: Annotation evaluation workflow. First, the annotation is represented in RDF. Then it is enriched. Lastly, we use the RDF-based machine learning to predict its quality.

therefore, we applied the following strategy: if any of the reviews stated the usefulness of an annotation as `usefulness-useful`, we selected that review, giving more weight to a potentially useful annotation. If not, we selected the usefulness value with the single highest frequency. When there were multiple reviews with the highest frequency, we removed the annotation as this happened in very few cases. Also we removed the reviewer information from the graph since that information would not be present for future (un-reviewed) annotations which we want to automatically assess.

Enrichment of the annotations is done since RDF graph kernels can easily use additional information to make predictions. The properties related to the artwork, the creator of the artwork and the annotation itself are relevant to be enriched. Unfortunately, to the best of our knowledge, there were no publicly accessible knowledge repositories related to artworks. We extend the creator data using the Union List of Artist Names (ULAN) and DBPedia, and annotation data with DBPedia, Flickr and Wikipedia as described in detail in Chapter 2.

For the third step we use Support Vector Machines and the Weisfeiler-Lehman graph kernel to estimate the quality of the annotation, exploiting the information provided in the enriched graph and using a set of previously evaluated (and enriched) annotations.

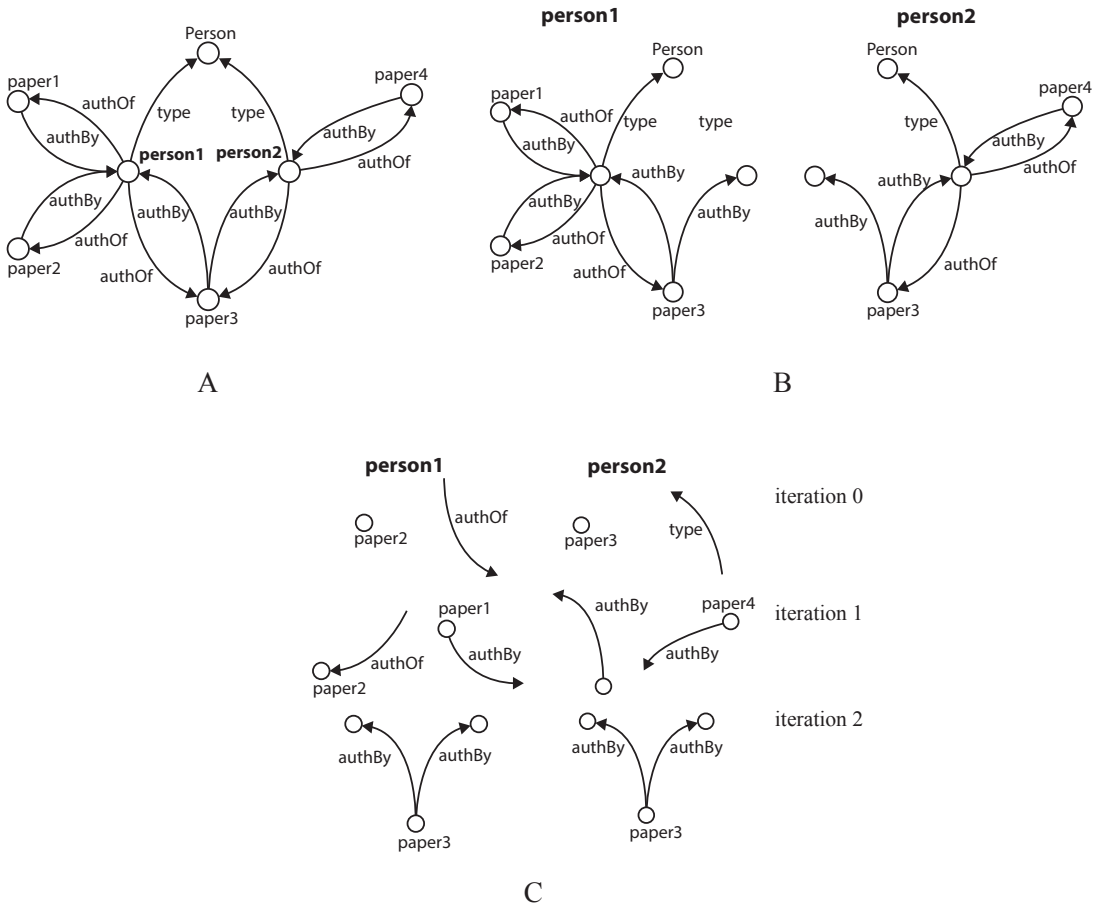

FIGURE 8.2: Example RDF graph (A), with two subgraphs of depth 2 (B) and examples of extracted features (C).

8.2.2 Machine learning with graph kernels for RDF

In a typical machine learning classification task, one tries to predict a class for a set of instances. Each instance is represented by a feature vector: a list of properties of that instance. This approach fits well to the scenario where the dataset is a table in a database, and each instance is a row. But it does not easily translate to RDF graphs. For example, consider the simple RDF graph given in Figure 8.2A. Suppose we want to predict a property of things that are Persons. Then our instances are the two nodes person1 and person2. It is not immediately obvious what the features of person1 and person2 are.

Machine learning for RDF data using graph kernels was introduced in [68] as a way to deal with this issue by using structural patterns of the RDF graph as input for kernel based learning algorithms [89, 91]. For each instance we consider the subgraph around that instance (up to a certain depth) as its 'features', see Figure 8.2B. For these subgraphs structural properties are computed as something that is called a 'kernel', which is essentially a similarity function between objects, for instance, between subgraphs of

an RDF graph. This kernel is used as the input data for a learning algorithm. The main advantage of using graph kernels for learning from RDF, compared to other techniques, is that it is a generically applicable and flexible approach [84]. Little knowledge of the dataset is required to use these methods and it allows for easy integration of additional knowledge into the learning process, by simply adding triples to the RDF graph.

In this chapter we will use the Weisfeiler-Lehman graph kernel for RDF (WLRDF). For each instance, the WLRDF kernel efficiently computes subtree patterns as features, in a number of iterations, where each iteration computes more complex patterns. These patterns are illustrated in Figure 8.2C. Typically, the features that are considered by a kernel are computed implicitly. However, subtree features of the WLRDF kernel are computed explicitly and we can therefore inspect which subtree patterns are important in the learning process.

As our learning algorithm, we use Support Vector Machine described in Chapter 2. In the machine learning step in our workflow, the instances that we use are annotations, i.e. nodes that are of type Annotation. For each annotation a subgraph is extracted up to a specified depth. From the semantic model with enriched annotations in Chapter 2, Figure 2.2, we can see that larger depths leads to the inclusion of more levels of information in the graph. The WLRDF kernel is computed using these subgraphs and then used to train a Support Vector Machine on labelled (in terms of quality) annotations. This Support Vector Machine is then used to predict the annotation quality of unseen annotations.

8.3 Experimental setup

We apply our approach on the `Steve.Museum` dataset. For our experiments we have `usefulness-useful`, `usefulness-not_useful`, `problematic` and `judgement` categories and removed both the `comments` category and annotations which were not evaluated.

The experiments were run on depth 1 (including annotation properties), depth 2 (additionally including annotator and artwork properties) and depth 3 (additionally including properties from the linked datasets). On each depth we created 10 subsets of the graph and performed a 5-fold cross-validation, optimizing the C-parameter of the Support Vector Machine in each fold, using again 5-fold cross-validation. The parameter C controls the trade off between errors of the SVM on training data and margin maximization. The number of iterations parameter, h, for the WLRDF kernel was fixed to the depth $\times 2$. This parameter can be optimized, but this has relatively little impact, since the higher iterations include the lower iterations. Subsets were created by taking a random sample

of annotations in the `usefulness-useful` category of size equal to the other categories combined and taking all annotations from those categories. Each subset contains approximately 9000 annotations. For each depth and subset we calculated the accuracy, precision, recall and F-measure for the categories combined and individually[1].

8.4 Discussion

In this section we discuss our experimental results. We provide both quantitative and qualitatively analysis of important features for predictions.

8.4.1 Comparison of accuracy, precision and recall for predictions at different depths

We compare the accuracy, F-measure, precision and recall for predicting four different categories (`usefulness-useful`, `usefulness-not_useful`, `judgment`, `problematic`) at three different depths of the graph and present the results in Table 8.1. The features for the graph which were included at different depths are described in Section 8.3. We repeated the experiment for predicting two types of review categories (`usefulness-useful` and `usefulness-not_useful`) and found that the results are comparable to the ones mentioned in Table 8.1, while the overall F-measure is higher, with 0.76 for every depth. This is to be expected since the two classes that were hard to predict are not included. The best overall results were achieved under the depth 2 setting. The `judgement` class is very hard to predict, as we can see from the very low precision, recall and F-measures.

8.4.2 Comparison of relevant graph features at different depths

The multi-class Support Vector Machine implementation in LibLINEAR computes a Support Vector Machine for each class, which can be used to identify the important graph features for each class. Thus, we trained a Support Vector Machine for the first of our 10 four-class subsets. A manual analysis of these important features (those with the highest weight) for the different classes at different depths shows some interesting results. We will not mention the results for the `judgement` class, since it was predicted very poorly.

At depth 1, the `useful` class has a large number of specific date strings, e.g. "2007-07-18T00:22:04", as important features. However, the `not-useful` class is recognized

[1]Our code can be found at `https://github.com/anottamkandath/Datasets/tree/master/Chapter8`

Table 8.1: Comparison of Results from Predictions Using the WLRDF Kernel at Different Depths

Depth	Prediction class	Avg. Accuracy	Precision	Recall	F-measure
1	Usefulness-useful		0.75	0.78	0.76
	Usefulness-not_useful		0.74	0.74	0.74
	Judgement		0.00	0.00	0.00
	Problematic		0.68	0.25	0.37
	All classes	0.75	0.54	0.44	0.47
2	Usefulness-useful		0.77	0.77	0.77
	Usefulness-not_useful		0.74	0.75	0.75
	Judgement		0.30	0.04	0.07
	Problematic		0.64	0.34	0.45
	All classes	0.75	0.61	0.48	0.51
3	Usefulness-useful		0.77	0.76	0.77
	Usefulness-not_useful		0.74	0.76	0.75
	Judgement		0.05	0.01	0.01
	Problematic		0.64	0.32	0.42
	All classes	0.75	0.55	0.46	0.49

by features pointing to the artwork that is annotated, such as `oac:hasTarget->http://purl.org/artwork/1043`. The `problematic` class has important features similar to the `useful` class.

Graph features containing the type of object and the annotation itself are almost exclusively the most important features for identifying `useful` annotations at depth 2 and 3. In contrast, the types of features that are used in classifying `not-useful` annotations are more diverse. They include graph features with the material used in the artwork or information about the annotators. For example, a set of important features has the graph pattern that includes the information that the annotator has "Intermediate" experience. The `problematic` class at depth 2 and 3 is recognized with very specific features, like date strings, that are not as general as for the other two classes.

8.5 Conclusion

In this chapter we presented a workflow to convert datasets in the cultural heritage domain to RDF and to enrich the datasets to be used for predictions of annotation quality using RDF graph kernels. We have provided both a qualitative and quantitative analysis of the results and have shown that RDF kernels are beneficial in making predictions about quality.

From our experiments it can be seen that employing RDF graph kernels helps in predicting classes of annotations with a overall best accuracy of 75%, which is a good rate of

acceptance. The individual class measures of accuracy, precision, recall and F-measure for the classes of `judgement` and `problematic` are not useful since the percentage of their classes were too small to perform a good training and thus they were predicted badly.

We also identified which features are relevant at different depths to make the predictions per category and provided an analysis. The features which are relevant to predict a certain class of quality are useful to design annotation tasks in the future. If a particular creator is selected as a relevant feature and if the majority of annotations by different users to an artwork belonging to that creator tend to be evaluated mostly as `usefulness-not_useful`, then it might indicate that the annotation task is difficult for that particular artwork. Similarily for different datasets such in-depth analysis helps to re-design the annotation tasks to obtain better quality from the crowds.

The approach of using graph kernels for RDF is flexible as additional information can easily be added to the learning process by extending the RDF graph. However, in the `Steve.Museum` dataset some node labels provide very specific information, which is not beneficial for generalization. For example, the annotations are timestamped with exact times in seconds, whereas the day of the week might be more informative. Some (light) graph pre-processing can help to alleviate these issues, without hindering the flexibility and extensibility of the approach.

Chapter 9

Conclusion

9.1 Research questions revisited

In this thesis we have described a number of methods for (semi-)automatic assessment of trust of annotations and annotators. The main research question of determining trust of cultural artefacts had been split into different research questions in the introduction and in this section we revisit each of the research questions and subsequently discuss the conclusions and future directions of our work.

- How can we model reputation of annotators from the crowd and quality of annotations regarding cultural heritage artefacts?

  In Chapter 3 we presented the workflow and algorithm for annotators providing annotations for cultural artefacts. The workflow demonstrates how the annotators registered with the cultural heritage institutions provide annotations and how artefacts are recommended to annotators based on their proven expertise. We used Semantic Web ontologies to model annotator profiles, their expertise and quality of annotations. We employed mechanisms for evaluating quality of provided annotations, and constantly managed and updated the trust, reputation and expertise information of registered annotators using subjective logic and semantic similarity measures. The evaluation of our model on the `Steve.Museum` dataset has shown the relevance of semantic similarity measures to compute trust. Thus we addressed the question of modelling the reputation of annotators and quality of annotations for cultural heritage artefacts.

- How can different techniques from probabilistic modelling be used to model trust?

In Chapter 4 we showed the potential for employing subjective logic as a basis for reasoning on Web and Semantic Web data. This work extended the results from Chapter 3, but explored different techniques to determine which operators in subjective logic provide better results for quality predictions and compared their performance. We also experimented with different semantic similarity measures such as deterministic and probabilistic measures to understand their effects on quality predictions. Part of this work is based on previously mentioned practical applications that show the usefulness of it, and we provided theoretical foundations for it. We also presented a technique to evaluate annotations when ground truth data is not available. Thus we answered the research question of employing different techniques in probabilistic modelling to model and determine trust in detail.

- How can demographics of annotator and provenance techniques be employed to evaluate quality of annotations?

  In Chapter 5 we investigated the correlation between different properties of the annotator and the quality of annotation. We also formed stereotypes of annotators based on their profile information and used them for prediction of quality. However, in the *Waisda?* dataset, we found that annotator stereotypes were not useful to discriminate annotator reputation, although we found a correlation between individual demographics (age, gender, etc.) and reputation. Moreover, we showed how to use the FOAF ontology to represent both annotator profiles and stereotypes. Additionally, we proposed and evaluated procedures for computing trust assessments based on reputation, for computing trust assessments based on provenance information, and for combining these two types of assessments. We showed that using reputation for trust assessment is simple, computationally light and accurate. We also showed the potential of provenance-based trust assessments: these can be at least as accurate as reputation-based methods and can be used to overcome the limitations of reputation-based approaches (at least within a tagging environment). For *Waisda?* dataset the combination of the two methods was more powerful than each of the two alone. Thus different techniques using annotator demographics, annotator reputation and provenance have been designed and evaluated to assess quality of cultural heritage artefacts and reputation of annotators.

- How can efficient techniques be developed for assessing quality of annotations?

  In the previous chapters, we presented algorithms for semi-automatically evaluating the quality of annotations by using subjective logic and semantic similarity to learn a model from a limited set of annotations evaluated by an institution. In Chapter 6, we introduced two extensions of this algorithm. The first extension

made use of semantic similarity to cluster the set of evaluated annotations at our disposal (training set) and hence improve the computational efficiency of the algorithm. The second extension regards the possibility to adapt the algorithm to use provenance information instead of the annotator reputation as a basis for the trustworthiness estimations. We evaluated each algorithm on two different datasets of annotations from the cultural heritage domain. The algorithm based on annotator reputation satisfactorily allows us to estimate the annotation trustworthiness with an accuracy of about 80% in one case and 70% in the other one. Thus clustering effectively helps in increasing the efficiency of the first extension, and the use of provenance information actually allow us to compute accurate estimates of annotations trustworthiness.

- How can machine learning techniques be applied on annotation and annotator features to make predictions on annotation reputation and quality of annotations?

  In Chapter 7, we determined what is the impact of different annotator demographics such as age, gender, education, etc. and different properties of annotation and provenance of annotation process on quality of information. We used machine learning prediction techniques by providing features of annotator, annotation and provenance for training the algorithms. This chapter is built on the initial results we obtained from our work in Chapter 5. In this chapter we described in detail the process of conversion of crowdsourced data in cultural heritage to Linked Data, identification of relevant features which affect the quality of annotations and reputation of annotators, modelling annotator reputation and employing machine learning algorithms to perform predictions about annotation quality and annotator reputation. The annotation quality predictions were performed using different sets of features and achieved a best overall average accuracy of 84% on the `Steve.Museum` dataset. Also the properties of annotators together with properties of annotations enhance prediction capabilities of categories with smaller training samples. The predictions of annotator reputation did not achieve better results than a random classification (of 50%).

- How can semantic relations and graph properties be combined with machine learning techniques for computing quality of annotations?

  In Chapter 8, we extended on our work in Chapter 5. Instead of using independent properties as features for the machine learning algorithms, we used semantic relation graphs depicting the relation between different entities and then reasoned on these graphs to determine quality of annotations. Thus we also utilised the semantic relationships between these entities and exploited them for machine learning

predictions. We presented a workflow to convert datasets in the cultural heritage domain to Resource Description Framework (RDF) and to enrich the datasets to be used for predictions of annotation quality using RDF graph kernels. We have provided both a qualitative and quantitative analysis of the results and have shown that RDF kernels are beneficial in making predictions about quality. From our experiments on `Steve.Museum` dataset, it can be seen that employing RDF graph kernels help in predicting classes of annotations with an overall best accuracy of 75%. We could also identify which features are relevant at different depths to make the predictions per category. Thus semantic relations and graph properties in combination with machine learning techniques help in computing quality of annotations.

Trust is a subjective concept and one should accept the fact that in certain cases there are no distinct answers and leave more room to uncertainty in opinions. With the growth of information on the Web and with active contributions from online users, it becomes necessary to devise algorithms to automate the evaluation of the quality of the contributed information. Our methods have proven to evaluate annotator contributed tags in the cultural heritage domain with relatively high accuracy. We will aim at further reducing the need for evaluated annotations to bootstrap our system, to reduce the burden on cultural heritage institutions in this process, and also investigate methods for further increasing the accuracy of our algorithms. This can be vital for the cultural heritage institutions which do not have many resources in terms of labour or finances at their disposal and decide to rely on crowdsourcing platforms, as well as for many other institutions and evaluating trust of data on the Web in general.

9.2 Implications for future work

The main assumptions in our approaches have been the availability of evaluated annotations for training our models to simulate the assessment techniques of cultural heritage institutions. This is because currently these institutions have their own methods for evaluation and their own standards for quality. We propose following directions of future work.

A future work would be to train our quality models with different cultural heritage institutions since these institutions gather different metadata and features for training the algorithms would vary and determining quality would depend on different factors. Thus the lessons learnt from one institution can be applied to others and would help us build a generic procedure for determining trust.

Crowdsourcing frameworks such as Crowdflower and Amazon's Mechanical turk can be exploited for obtaining annotations. Users perform tasks in such frameworks for monetary incentives and cultural heritage institutions had from the past always focused on community driven methods for enriching their collections. Thus it would be a good study to observe the performance of different incentives for annotating cultural heritage artefacts. Some of the lessons learnt from our work can be applied while designing tasks in such crowdsourcing frameworks. We had shown the importance of gathering user data and provenance information for annotation tasks, and thus tasks can be defined in a way to gather more details about the user and also gather provenance information. Also the nature of tasks provided on such platforms can be more high-level and would help in initial classification of artefacts before providing them for annotation tasks to experts. Our techniques of evaluations using partial evidence can be utilised in such platforms, where multiple users agree on the results of a single task to evaluate the quality. More investigations can be performed regarding the user experience in such platforms and how it compares to the current techniques and eventually understand which techniques help result in better quality annotations.

Another interesting direction would be employing image recognition software to identify objects in the artefacts. This would mean that the software is able to identify different objects from various eras of art and diverse artist styles. If a confidence score regarding the object identification is obtained, we can combine it with trust algorithms to determine with more certainty about the presence of an object in a painting. The drawback of this approach is that it is only applicable for annotations that describe visually identifiable objects.

There exists many ontologies on the Semantic Web and different techniques to compute semantic similarity. In the future we will investigate the better integration of semantic similarity measures in subjective logic, to make it more standardised, and possibly provide best practices that help choosing the right ontologies and semantic similarity measures for mapping based on a given set of requirements.

We aim to further investigate techniques to identify the relevant provenance features for our algorithms, and the possibility of automatically extracting provenance patterns usable for trust assessment, to automate, optimise and adapt the process to other case studies and domains.

Based on the meta-data and a subset of annotations provided by annotators, it will be possible to automatically suggest annotations for artefacts and have them verified by the crowd. Our work uses semi-automated techniques. As future work we would like to move to more automated techniques.

If in the future the cultural heritage institutions decide to link their collections on the Web, there can be common policies regarding the annotation gathering and storage process and there would be common expectations of quality of annotations and reputations of annotators.

We observed that there is issue of variability regarding quality of annotations between reviewers. If an annotation has been given only one review, it does not necessarily mean that annotation belongs to that particular review category. It could have been that there were insufficient resources to obtain more reviews for that annotation. Although institutions do provide guidelines to review annotations, trust is a subjective measure and different reviewers assign different quality measures to the same annotation even though they use the same guidelines. More research can be performed to analyse and solve this issue for employees within an institution and later these can be mapped for solving quality variability issues between different institutions.

Cultural heritage institutions along with researchers in psychology can explore about human behaviour and their relations with respect to trust. The research should also incorporate the fact that humans in the real world perceive trust differently that in a digital environment. Cultural heritage institutions can investigate more about better ways to develop and maintain communities of users since performance of users depends on tasks which provide them a greater sense of purpose, once their basic needs are satisfied. Thus development efforts for creating such platforms and how different optimisations in the platform can affect the performance of annotators can be investigated. The ultimate goal it to enable the public to enhance collections while engaging with and exploring them.

Bibliography

[1] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Analyzing user modeling on Twitter for personalized news recommendations. In *Proceedings of the 19th Conference on User Modeling, Adaption and Personalization - UMAP 2011*, volume 6787 of *Lecture Notes in Computer Science*, pages 1–12. Springer.

[2] Altintas, I., Anand, M. K., Crawl, D., Bowers, S., Belloum, A., Missier, P., Ludäscher, B., Goble, C. A., and Sloot, P. M. A. (2010). Understanding collaborative studies through interoperable workflow provenance. In *Proceedings of the 2nd International Conference on Provenance and Annotation of Data and Processes*, International Provenance and Annotation Workshop 2010, pages 42–58. Springer.

[3] Aroyo, L. and Welty, C. (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM.

[4] Artz, D. and Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2):131–197.

[5] Berners-Lee, T. and Fischetti, M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper San Francisco.

[6] Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., and Weitzner, D. J. (2006). A framework for web science. *Found. Trends Web Sci.*, pages 1–130.

[7] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, pages 34–43.

[8] Bizer, C. and Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics*, 7(1):1–10.

[9] Breslin, J. G., Bojars, U., Aleman-meza, B., Boley, H., Nixon, L. J., Polleres, A., and Zhdanova, A. V. (2007). Finding experts using internet-based discussions in online communities and associated social networks. In *Finding Experts on the Web with Semantics*, pages 38–47. CEUR-WS.org.

[10] Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

[11] Burnett, C., Normal, T. J., and Sycara, K. (2010). Bootstrapping trust evaluations through stereotypes. In *Autonomous Agents and Multi Agent Systems*, pages 241–248. IFAAMAS.

[12] Card, S., Moran, T. P., and Newell, A. (1983). *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates.

[13] Carroll, J., Bizer, C., Hayes, P., and Stickler, P. (2005). Named graphs, provenance and trust. In *World Wide Web 2005*, pages 613–622. ACM.

[14] Castelfranchi, C. and Falcone, R. (1998). Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *International Conference on Multi Agent Systems 1998*, pages 72–79. IEEE Computer Society.

[15] Ceolin, D. (2014). *Trusting Semi-structured Web Data*. PhD thesis, VU University Amsterdam, The Netherlands.

[16] Ceolin, D., Groth, P., and van Hage, W. R. (2010a). Calculating the trust of event descriptions using provenance. In *Semantic Web and Provenance Management 2010*. CEUR-WS.org.

[17] Ceolin, D., van Hage, W., and Fokkink, W. (2010b). A trust model to estimate the quality of annotations using the Web. In *Proceedings of the 2010 Web Science Conference*. ACM.

[18] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *Advanced Computing Machinery Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

[19] Chen, J., Huang, H., Tian, S., and Qu, Y. (2009). Feature selection for text classification with naive Bayes. *Expert Systems Applications*, 36(3):5432–5435.

[20] Cilibrasi, R. and Vitányi, P. M. B. (2006). Automatic meaning discovery using Google. In *Kolmogorov Complexity and Applications*, Dagstuhl Seminar Proceedings.

[21] Cilibrasi, R. L. and Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.

[22] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[23] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2007). SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Intelligent User Interfaces*, pages 32–41. ACM.

[24] Damme, C. V. and Coenen, T. (2008). Quality metrics for tags of broad folksonomies. In *Proceedings of the 2008 International Conference on Semantic Systems*, pages 118–125. Journal of Universal Computer Science.

[25] De la Calzada, G. and Dekhtyar, A. (2010). On measuring the quality of Wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility*, pages 11–18. ACM.

[26] de Vries, G. K. D. (2013). A fast approximation of the Weisfeiler-Lehman graph kernel for RDF data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 606–621. LNCS Springer.

[27] Demartini, G. (2007). Finding Experts Using Wikipedia. In *Proceedings of the 2nd International International Semantic Web Conference +Asian Semantic Web Conference Workshop on Finding Experts on the Web with Semantics*, pages 33–41. CEUR-WS.org.

[28] Ebden, M., Huynh, T. D., Moreau, L., Ramchurn, S., and Roberts, S. (2012). Network analysis on provenance graphs from a crowdsourcing application. In *Proceedings of the 4th International Conference on Provenance and Annotation of Data and Processes*, pages 168–182. Springer.

[29] Ellis, A., Gluckman, D., Cooper, A., and Greg, A. (2012). Your paintings: A nation's oil paintings go online, tagged by the public. In *Proceedings of Museums and the Web 2012*.

[30] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

[31] Fink, D. (1995). A Compendium of Conjugate Priors. Technical report, Cornell University.

[32] G. Begelman, P. K. and Smadja, F. (2006). Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the Collaborative Web Tagging Workshop*, pages 15–33.

[33] Gambetta, D. (1988). *Can We Trust Trust?* Basil Blackwell.

[34] Gamble, M. and Goble, C. (2011). Quality, trust, and utility of scientific data on the web: Towards a joint model. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 15:1–15:8. ACM.

[35] Georgescu, M. and Zhu, X. (2014). Aggregation of crowdsourced labels based on worker history. In *Proceedings of the 4th Conference on Web Intelligence, Mining and Semantics*, pages 1–11. ACM.

[36] Glass, G. V. and Hopkins, K. D. (1995). *Statistical Methods in Education and Psychology*. Allyn & Bacon.

[37] Golbeck, J. and Hendler, J. (2004a). *Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks*. Springer Berlin Heidelberg.

[38] Golbeck, J. and Hendler, J. (2004b). Inferring reputation on the semantic web.

[39] Gower, J.C. and Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistics Society*, 18(1):54–64.

[40] Grosse-Bolting, G., Nishioka, C., and Scherp, A. (2015). Generic process for extracting user profiles from social media using hierarchical knowledge bases. In *Proceedings of the 9th Conference on Semantic Computing*, pages 197–200. IEEE Computer Society.

[41] Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

[42] Hartig, O. and Zhao, J. (2009). Using web data provenance for quality assessment. In *Semantic Web and Provenance Management 2009*, pages 26–31. CEUR-WS.org.

[43] Hassan-Montero, Y. and Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*. ACL.

[44] Heath, T. and Motta, E. (2008). The Hoonoh ontology for describing trust relationships in information seeking. In *Personal Identification and Collaborations: Knowledge Media and Extraction 2008*. CEUR-WS.org.

[45] Hennicke, S., Olensky, M., de Boer, V., Isaac, A., and Wielemaker, J. (2011). A data model for cross-domain data representation. the "Europeana data model" in the case of archival and museum data. In *Proceedings des 12. Internationalen Symposiums der Informationswissenschaft*, pages 136–147. Verlag Werner Hulsbusch.

[46] Hildebrand, M., Brinkerink, M., Gligorov, R., Steenbergen, M. V., Huijkman, J., and Oomen, J. (2013). Waisda?: Video labeling game. In *Proceedings of Advanced Computing Machinery Multimedia 2013*, pages 823–826. ACM.

[47] Hirth, M., Hossfeld, T., and Tran-Gia, P. (2011). Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In *Innovative Mobile and Internet Services in Ubiquitous Computing*, pages 316–321. IEEE Computer Society.

[48] Hollink, L., Malaisé, V., and Schreiber, G. (2010). Thesaurus enrichment for query expansion in audiovisual archives. *Multimedia Tools Applications*, 49(1):235–257.

[49] Inel, O., Aroyo, L., Welty, C., and Sips, R.-J. (2013). Domain-independent quality measures for crowd truth disagreement. *Journal of Detection, Representation, and Exploitation of Events in the Semantic Web*, pages 2–13.

[50] Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., and Sips, R.-J. (2014). Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *International Semantic Web Conference 2014*, volume 8797 of *Lecture Notes in Computer Science*, pages 486–504. Springer.

[51] Javanmardi, S., Lopes, C., and Baldi, P. (2010). Modeling user reputation in wikis. *Statistical Analysis Data Mining*, 3(2):126–139.

[52] John, G. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann.

[53] Jøsang, A. (2001). A logic for uncertain probabilities. *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212.

[54] Jøsang, A. (2002). The consensus operator for combining beliefs. *Artificial Intelligence Journal*, 142:157–170.

[55] Jøsang, A., Daniel, M., and Vannoorenberghe, P. (2003). Strategies for combining conflicting dogmatic beliefs. In *Proceedings of the 6th IEEE International Conference on Information Fusion*, pages 1133–1140. IEEE Computer Society.

[56] Jøsang, A., Diaz, J., and Rifqi, M. (2010). Cumulative and averaging fusion of beliefs. *Journal of Information Fusion*, 11(2):192–200.

[57] Jøsang, A., Marsh, S., and Pope, S. (2006). Exploring different types of trust propagation. In *Proceedings of the 4th International Conference on Trust Management*, pages 179–192. Springer.

[58] Jøsang, A. and McAnally, D. (2005). Multiplication and comultiplication of beliefs. *International Journal of Approximate Reasoning*, 38(1):19–51.

[59] Kaplan, L., Chakraborty, S., and Bisdikian, C. (2012). Subjective logic with uncertain partial observations. In *Proceedings of the 15th IEEE International Conference on Information Fusion*. IEEE Computer Society.

[60] Kassing, S., Oosterman, J., Bozzon, A., and Houben, G.-J. (2015). Locating domain-specific contents and experts on social bookmarking communities. In *Proceedings of 30th Symposium on Applied Computing*. ACM.

[61] Kazai, G., Kamps, J., and Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Conference on Information and Knowledge Management*, pages 2583–2586. ACM.

[62] Kononenko, I. (1992). Naive Bayesian classifier and continuous attributes. *Informatica*, 16(1):1–8.

[63] Lease, M., Yilmaz, E., Sorokin, A., and Hester, V., editors (2011). *Proceedings of the 2nd Workshop on Crowdsourcing for Information Retrieval*. ACM.

[64] Leyssen, M. H. R., Traub, M. C., van Ossenbruggen, J. R., and Hardman, L. (2012). Is It A Bird Or Is It A Crow? The Influence Of Presented Tags On Image Tagging By Non-Expert Users. Technical Report INS-1202, CWI.

[65] Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann Publishers.

[66] Liu, X., Datta, A., Rzadca, K., and Lim, E.-P. (2009). StereoTrust: A group based personalized trust model. In *Conference on Information and Knowledge Management*, pages 7–16. ACM.

[67] Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics 2002*, pages 63–70. Association for Computational Linguistics.

[68] Lösch, U., Bloehdorn, S., and Rettinger, A. (2012). Graph kernels for RDF data. In *Extended Semantic Web Conference*, pages 134–148. LNCS Springer.

[69] Masum, H. and Tovey, M., editors (2012). *The Reputation Society*. MIT Press.

[70] McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceedings of Association for the Advancement of Artificial Intelligence-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press.

[71] Medelyan, O., Frank, E., and Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327. Association for Computational Linguistics.

[72] Miller, G. A. (1995). Wordnet: a lexical database for English. *Commununications of the ACM*, 38(11):39–41.

[73] Milne, D. and Witten, I. (2008a). Learning to link with Wikipedia. In *Conference on Information and Knowledge Management*, pages 509–518. ACM.

[74] Milne, D. and Witten, I. H. (2008b). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30. AAAI Press.

[75] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and den Bussche, J. V. (2011). The open provenance model core specification (v1.1). *Future Generations Computer Systems*, 27(6):743–756.

[76] O'Hara, K. (2012). A General Definition of Trust. Technical report, University of Southampton.

[77] Olmedilla, D., Rana, O. F., Matthews, B., and Nejdl, W. (2005). Security and trust issues in semantic grids. In *In Proceedings of the Dagsthul Seminar, Semantic Grid: The Convergence of Technologies*, page 05271.

[78] Oomen, J. and Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: Opportunities and challenges. In *Proceedings of the 5th International Conference on Communities and Technologies*, pages 138–149. ACM.

[79] Pantola, A. V., Pancho-Festin, S., and Salvador, F. (2010). Rating the raters: A reputation system for wiki-like domains. In *Security of Information and Networks 2010*, pages 71–80. ACM.

[80] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175.

[81] Prasad, T. K., Anantharam, P., Henson, C. A., and Sheth, A. P. (2014). Comparative trust management with applications: Bayesian approaches emphasis. *Future Generation Computer Systems*, pages 182–199.

[82] Rajbhandari, S., Rana, O. F., and Wootten, I. (2008). A fuzzy model for calculating workflow trust using provenance data. In *Mardi Gras 2008*, pages 1–8. ACM.

[83] Rajbhandari, S., Wootten, I., Ali, A. S., and Rana, O. F. (2006). Evaluating provenance-based trust for scientific workflows. In *Cloud Cluster and GRID Computing 2006*, pages 365–372. IEEE Computer Society.

[84] Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., and Fanizzi, N. (2012). Mining the semantic web—statistical learning for next generation knowledge bases. *Data Mining Knowledge Discovery*, 24(3):613–662.

[85] Ridge, M. (2013). From tagging to theorizing: Deepening engagement with cultural heritage through crowdsourcing. *Curator: The Museum Journal*, 56(4):435–450.

[86] Ridge, M. (2014). Introduction. In Ridge, M., editor, *Crowdsourcing Our Cultural Heritage*, Digital Research in the Arts and Humanities. Ashgate.

[87] Sabater, J. and Sierra, C. (2005). Review on computational trust and reputation models. *Artificial Intelligence Review*, 24:33–60.

[88] Sanderson, R., Ciccarese, P., de Sompel, H. V., Clark, T., Cole, T., Hunter, J., and Fraistat, N. (2012). Open Annotation Core Data Model. Technical report, W3C Community.

[89] Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

[90] Secord, A. (1994). Corresponding interests: Artisans and gentlemen in nineteenth-century natural history. *The British Journal for the History of Science*, 27(4):383–408.

[91] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

[92] Shervashidze, N. and Borgwardt, K. M. (2009). Fast subtree kernels on graphs. In *Advances in Neural Information Processing Systems*, pages 1660–1668. Curran Associates, Inc.

[93] Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. (2011). Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561.

[94] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

[95] Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Social Computing 2010*, pages 177–184. IEEE Computer Society.

[96] Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Anchor.

[97] Trant, J. (2009). Tagging, folksonomy and art museums: Early experiments and ongoing research. *Journal of Digital Informatics*, 10(1).

[98] Ushioda, A. (1996). Hierarchical clustering of words and application to NLP tasks. In *International Conference on Computational Linguistics*, pages 28–41. ACL.

[99] Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. (2014). Community-based Bayesian aggregation models for crowdsourcing. In *World Wide Web*, pages 155–164. ACM.

[100] von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326. ACM.

[101] Vossen, P., Hofmann, K., de Rijke, M., Sang, E. T. K., and Deschacht, K. (2007). The Cornetto database: Architecture and user-scenarios. In *Proceedings of 7th Dutch-Belgian Information Retrieval Workshop*, pages 89–96.

[102] Wang, S. and Iwaihara, M. (2011). Quality evaluation of Wikipedia articles through edit history and editor groups. In *Proceedings of the 13th Asia-Pacific Web Conference*, pages 188–199. LNCS Springer-Verlag.

[103] Wang, Y., Wang, S., Stash, N., Aroyo, L., and Schreiber, G. (2010). Enhancing content-based recommendation with the task model of classification. In *Knowledge Engineering and Management by the Masses*, volume 6317 of *Lecture Notes in Computer Science*, pages 431–440. Springer.

[104] Warncke-Wang, M., Cosley, D., and Riedl, J. (2013). Tell me more: An actionable quality model for Wikipedia. In *Proceedings of 9th Symposium on Open Collaboration*, pages 1–10. ACM.

[105] Weisfeiler, F. and Lehman., A. A. (1968). A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia*, 9:12–16.

[106] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

[107] Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

[108] Zaihrayeu, I., da Silva, P., and McGuinness, D. L. (2005). IWTrust: Improving user trust in answers from the Web. In *International Conference of Trust Management, iTrust2005*, pages 384–392. LNCS Springer.

[109] Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., and McGuinness, D. L. (2006). Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*, page 8. ACM.

[110] Zhou, Y., Cong, G., Cui, B., Jensen, C. S., and Yao, J. (2009). Routing questions to the right users in online communities. In *International Conference on Data Engineering*. IEEE Computer Society.

Summary

Cultural and heritage preserving organisations such as museums are rapidly digitising their collections, and at the same time migrating digitised collections to the Web. Through the Web, these institutions can reach large masses of people, with intentions varying from increasing visibility to acquiring user-generated content. To facilitate archiving and retrieval operations on the Web, collections must be described by high-quality annotations that cover physical properties (e.g. dimensions, material), provenance (e.g. creator, previous owners) and subject matter (e.g. what is represented) of the artworks. The annotations later becomes the metadata for the artworks in the institutions. Cultural heritage institutions employ professionals, mostly art historians, to provide high-quality annotations about art-historical properties. They are trained and follow strict guidelines on how to correctly and qualitatively annotate artefacts; but their effectiveness is hindered by different factors such as the size of museum collections (which can be in the order of millions of artworks), temporal and monetary constraints, and a lack of domain expertise on some of the subject matter of artworks.

For this reason, many cultural heritage institutions have opened up their archives to ask the masses to help them in tagging or annotating their artefacts. In earlier years it was feasible for employees at the cultural heritage institutions to manually assess the quality of tags entered by external annotators, since there were relatively few contributions from Web users. However, with the growth of the Web, the amount of data has become too large to be accurately dealt with by experts at the disposal of these institutions within a reasonable time. Cultural heritage institutions need the annotations to be trustworthy in order to maintain their authoritative reputation. In this thesis, our challenge is to build algorithms which help to predict quality of annotations and reputation of annotators.

Trust is a subjective phenomenon and humans use the concept of trust in various situations. An annotation considered trustworthy by one institution may not be considered in the same manner by another one. Thus it is important to understand how trust of information is defined and understood by different institutions.

In this thesis we addresses various research questions. The first research question is *How can we model reputation of annotators from the crowd and quality of annotations regarding cultural heritage artefacts?*. In Chapter 3 we investigated techniques to model and represent the reputation of annotators and the quality of annotations using subjective logic and semantic similarity measures. We proposed a workflow which can be employed by cultural heritage institutions to evaluate reputation of annotators and quality of provided annotations. This chapter served as an introductory point to the other chapters where we developed various techniques to model and determine trust. Chapter 4 addressed the second research question *How can different techniques in probabilistic modelling be used to model trust?*. We discussed how different operators in subjective logic can be used to model opinions and compared their performance. This is helpful to model opinions when ground truth data is available and we need to make future predictions. In case ground truth data is not available, we discussed about partial evidence. Subjective logic operators can be tuned to model such instances where there is only partial evidence such as multiple agreements available. The third research question *How can demographics of annotator and provenance techniques be employed to evaluate the quality of annotations?* is addressed in Chapter 5 where we investigated the correlation between different properties of the annotator and the quality of annotation. We also formed stereotypes of annotators and used them for prediction of quality. Apart from annotator demographics we used details about how an annotation was created (provenance) to determine quality of annotations. We later combined the techniques of determining trust based on reputation with techniques employing provenance for determining trust and compared their performance. Chapter 6 discussed how we answered the question *How can efficient techniques be developed for assessing the quality of annotations?*. We were able to increase the efficiency of our trust computation algorithms by decreasing the computation time while maintaining or increasing the performance of our algorithms. We used machine learning clustering techniques to group semantically similar annotations provided by annotators on the Web about different artefacts and also employed clustering based on provenance information. Machine learning techniques were explored more in detail in Chapter 7 which dealt with the question *How can machine learning techniques be applied on annotation and annotator features to make predictions on annotator reputation and the quality of annotations?*. We determined the impact of different annotator demographics such as age, gender, education, etc., and different properties of annotations and provenance of the annotation process on the quality of information. We use machine learning prediction techniques by providing features of annotator, annotation and provenance for training the algorithms. Our last research question *How can semantic relations and graph properties be combined with machine learning techniques for computing the quality of annotations?* was answered in Chapter 8. Instead of using

independent properties as features for the machine learning algorithms, we built semantic relation graphs depicting the relation between different entities and then reasoned on these graphs to determine the quality of annotations. Thus instead of using only the available features regarding the annotator, annotation and annotation process, we also utilised the semantic relationships between these entities and exploit them for machine learning predictions.

Thus in this thesis we have described various techniques which will help professionals from cultural heritage institutions to assess the quality of annotations and reputation of annotators enriching their collections.

Samenvatting

Culturele en erfgoed bewarende instellingen zoals musea digitaliseren in rap tempo hun collecties, en migreren tegelijkertijd digitale collecties naar het Web. Via het Web kunnen deze organisaties grote aantallen mensen bereiken, met intenties die variëren van toename in zichtbaarheid tot het verkrijgen van inhoudelijke bijdragen. Om archivering en het terugvinden van kunstwerken op het Web te faciliteren, dienen collecties te worden voorzien van hoogkwalitatieve annotaties over fysieke eigenschappen (bijv. dimensies, materiaal), herkomst (bijv. maker, vorige eigenaars) en onderwerp (bijv. wat wordt gerepresenteerd) van kunstwerken. Ook kunnen dergelijke annotaties door instellingen worden gebruikt als metadata over hun kunstwerken. Dergelijke annotaties worden aangeleverd door professionele werknemers, meestal kunsthistorici. Zij zijn getraind om strikte richtlijnen te volgen wat betreft het correct en kwalitatief annoteren van kunstwerken; maar hun effectiviteit wordt gehinderd door verschillende factoren zoals de omvang van museale collecties (die in de orde van miljoenen kunstwerken kan zijn), beperkingen in tijd en geld, en een gebrek aan specifieke domeinkennis op sommige onderwerpen in kunstwerken. Daarom openen veel culturele instellingen hun digitale archieven voor het publiek en vragen ze om hulp bij het labelen en annoteren van hun kunstwerken. Voor de reputatie van de instellingen is het van groot belang dat alleen hoogkwalitatieve annotaties worden verwerkt. Aanvankelijk was het voor medewerkers van dergelijke instellingen doenlijk om handmatig de kwaliteit van extern aangeleverde annotaties te evalueren. Met de snelle groei van het Web is de hoeveelheid aangeleverde data echter te groot geworden om dergelijke evaluaties accuraat uit te voeren. De centrale onderzoeksvraag in dit proefschrift is het ontwikkelen van algoritmes die ondersteuning bieden bij het inschatten van de kwaliteit van zowel annotaties als de reputatie van annotators.

Betrouwbaarheid is een subjectief begrip, en mensen passen dit concept toe in verschillende situaties. Een annotatie die betrouwbaar is bevonden door een culturele instelling wordt mogelijk niet op dezelfde manier beschouwd door een andere instelling. Het is belangrijk om te begrijpen hoe betrouwbaarheid van informatie is gedefinieerd en wordt geïnterpreteerd door verschillende instellingen.

In dit proefschrift zijn verschillende onderzoeksvragen geadresseerd. De eerste vraag is: *Hoe kunnen de reputatie van annotators uit het publiek en de kwaliteit van annotaties van kunstwerken worden gemodelleerd?*. In Hoofdstuk 3 onderzochten we technieken om deze modellering uit te voeren met behulp van zogenaamde subjectieve logica en semantische similariteitsmetrieken, en poneerden we een productiestroom die kan worden toegepast bij culturele instellingen. Dit hoofdstuk diende als uitgangspunt voor de overige hoofdstukken, waarin verschillende technieken zijn ontwikkeld voor het modelleren en bepalen van betrouwbaarheid. Hoofdstuk 4 adresseerde de tweede onderzoeksvraag: *Hoe kunnen probabilistische technieken worden toegepast bij het modelleren van betrouwbaarheid?*. We onderzochten hoe verschillende operatoren in subjectieve logica kunnen worden gebruikt om opinies te modelleren, en vergeleken hun prestaties. Dit is toepasbaar wanneer reeds data over betrouwbaarheid van annotators en annotaties beschikbaar is en we voorspellingen dienen te maken betreffende nieuwe data. In sommige gevallen is geen absolute data over betrouwbaarheid beschikbaar, maar alleen zogenaamd gedeeltelijk bewijs, zoals overeenkomsten of verschillen tussen annotators. Operatoren in subjectieve logica kunnen worden aangepast om op basis van dergelijke informatie voorspellingen te doen. Hoofdstuk 5 adresseerde de derde onderzoeksvraag: *Hoe kan demografische informatie over annotators en informatie over herkomst worden gebruikt om de kwaliteit van annotaties te evalueren?*. We onderzochten de correlatie tussen verschillende eigenschappen van een annotator en de kwaliteit van zijn of haar annotaties. We formuleerden ook stereotypes voor annotators en gebruikten ze om de kwaliteit van annotaties te voorspellen. Ook informatie over demografische gegevens van annotators en hoe een annotatie is gecreëerd werden gebruikt bij het evalueren van annotaties. We combineerden tenslotte deze technieken en vergeleken de prestaties van de afzonderlijke technieken en hun combinaties. Hoofdstuk 6 adresseerde de vierde onderzoeksvraag: *Hoe kunnen efficiënte technieken worden ontwikkeld voor het inschatten van de kwaliteit van annotaties?*. We waren in staat om de effectiviteit van onze algoritmes te verhogen door de berekeningstijd te verkorten zonder verlies van nauwkeurigheid. We gebruikten clusteringtechnieken, die in de context van zelflerende systemen zijn ontwikkeld, om semantisch vergelijkbare annotaties van verschillende annotators over verschillende kunstwerken te groeperen. Ook gebruikten we clustering op basis van informatie over herkomst. Hoofdstuk 7 adresseerde de onderzoeksvraag: *Hoe kunnen technieken afkomstig van zelflerende systemen worden toegepast op basis van eigenschappen van annotators en annotaties om voorspellingen te maken over de reputatie van annotators en de kwaliteit van annotaties?*. We bepaalden het effect van verschillende demografische karakteristieken van annotators zoals leeftijd, geslacht, opleiding, etc., alsmede verschillende karakteristieken van annotaties en de ontstaansgeschiedenis van annotaties, op de kwaliteit van informatie. We gebruikten zelflerende voorspellingstechnieken door het aanleveren van eigenschappen van annotators, annotaties en herkomst om de algoritmes te trainen. Hoofdstuk

8 adresseerde de laatste onderzoeksvraag: *Hoe kunnen semantische relatiens en eigenschappen van grafen worden gecombineerd met zelflerende technieken om de kwaliteit van annotaties te berekenen?*. In plaats van onafhankelijke eigenschappen te gebruiken als invoer voor de zelflerende algoritmes, bouwden we grafen die semantische relaties tussen verschillende entiteiten weergeven, en redeneerden over deze grafen om voorspellingen te doen over de kwaliteit van annotaties.

Concluderend zijn in dit proefschrift verscheidene technieken beschreven die werknemers van culturele instellingen kunnen helpen bij het beoordelen van de kwaliteit van annotaties en de reputatie van annotators, in de context van de verrijking van hun gedigitaliseerde collecties.

SIKS Dissertation Series

====
2010
====

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter

2010-02 Ingo Wassink (UT)
Work flows in Life Science

2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents

2010-04 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI

2010-07 Wim Fikkert (UT)
Gesture interaction at a Distance

2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-09 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning

2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis

2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques

2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration

2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models

2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice

2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems

2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation

2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data

2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions

2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies

2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-27 Marten Voulon (UL)
Automatisch contracteren

2010-28 Arne Koopman (UU)
Characteristic Relational Patterns

2010-29 Stratos Idreos(CWI)
Database Cracking: Towards Auto-tuning Database Kernels

2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web

2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions

2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-37 Niels Lohmann (TUE)
Correctness of services and their composition

2010-38 Dirk Fahland (TUE)
From Scenarios to components

2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents

2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web

2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search

2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services

2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples

2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions

2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives

2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources

2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention

2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

====
2011
====

2011-01 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02 Nick Tinnemeier(UU)
Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

2011-03 Jan Martijn van der Werf (TUE)
Compositional Design and Verification of Component-Based Information Systems

2011-04 Hado van Hasselt (UU)
Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms

2011-05 Base van der Raadt (VU)
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

2011-06 Yiwen Wang (TUE)
Semantically-Enhanced Recommendations in Cultural Heritage

2011-07 Yujia Cao (UT)
Multimodal Information Presentation for High Load Human Computer Interaction

2011-08 Nieske Vergunst (UU)
BDI-based Generation of Robust Task-Oriented Dialogues

2011-09 Tim de Jong (OU)
Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT)
Cloud Content Contention

2011-11 Dhaval Vyas (UT)
Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE)
Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT)
Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA)
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM)
Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA)
Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU)
The Mind's Eye on Personal Profiles

2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU))
Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein(UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-29 Faisal Kamiran (TUE)
Discrimination-aware Classification

2011-30 Egon van den Broek (UT)
Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-31 Ludo Waltman (EUR)
Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-32 Nees-Jan van Eck (EUR)
Methodological Advances in Bibliometric Mapping of Science

2011-33 Tom van der Weide (UU)
Arguing to Motivate Decisions

2011-34 Paolo Turrini (UU)
Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-35 Maaike Harbers (UU)
Explaining Agent Behavior in Virtual Training

2011-36 Erik van der Spek (UU)
Experiments in serious game design: a cognitive approach

2011-37 Adriana Burlutiu (RUN)
Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

2011-38 Nyree Lemmens (UM)
Bee-inspired Distributed Optimization

2011-39 Joost Westra (UU)
Organizing Adaptation using Agents in Serious Games

2011-40 Viktor Clerc (VU)
Architectural Knowledge Management in Global Software Development

2011-41 Luan Ibraimi (UT)
Cryptographically Enforced Distributed Data Access Control

2011-42 Michal Sindlar (UU)
Explaining Behavior through Mental State Attribution

2011-43 Henk van der Schuur (UU)
Process Improvement through Software Operation Knowledge

2011-44 Boris Reuderink (UT)
Robust Brain-Computer Interfaces

2011-45 Herman Stehouwer (UvT)
Statistical Language Models for Alternative Sequence Selection

2011-46 Beibei Hu (TUD)
Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

2011-47 Azizi Bin Ab Aziz(VU)
Exploring Computational Models for Intelligent Support of Persons with Depression

2011-48 Mark Ter Maat (UT)
Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

2011-49 Andreea Niculescu (UT)
Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

====
2012
====

2012-01 Terry Kakeeto (UvT)
Relationship Marketing for SMEs in Uganda

2012-02 Muhammad Umair(VU)
Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

2012-03 Adam Vanya (VU)
Supporting Architecture Evolution by Mining Software Repositories

2012-04 Jurriaan Souer (UU)
Development of Content Management System-based Web Applications

2012-05 Marijn Plomp (UU)
Maturing Interorganisational Information Systems

2012-06 Wolfgang Reinhardt (OU)
Awareness Support for Knowledge Workers in Research Networks

2012-07 Rianne van Lambalgen (VU)
When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

2012-08 Gerben de Vries (UVA)
Kernel Methods for Vessel Trajectories

2012-09 Ricardo Neisse (UT)
Trust and Privacy Management Support for Context-Aware Service Platforms

2012-10 David Smits (TUE)
Towards a Generic Distributed Adaptive Hypermedia Environment

2012-11 J.C.B. Rantham Prabhakara (TUE)
Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

2012-12 Kees van der Sluijs (TUE)
Model Driven Design and Data Integration in Semantic Web Information Systems

2012-13 Suleman Shahid (UvT)
Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

2012-14 Evgeny Knutov(TUE)
Generic Adaptation Framework for Unifying Adaptive Web-based Systems

2012-15 Natalie van der Wal (VU)
Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.

2012-16 Fiemke Both (VU)
Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

2012-17 Amal Elgammal (UvT)
Towards a Comprehensive Framework for Business Process Compliance

2012-18 Eltjo Poort (VU)
Improving Solution Architecting Practices

2012-19 Helen Schonenberg (TUE)
What's Next? Operational Support for Business Process Execution

2012-20 Ali Bahramisharif (RUN)
Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

2012-21 Roberto Cornacchia (TUD)
Querying Sparse Matrices for Information Retrieval

2012-22 Thijs Vis (UvT)
Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

2012-23 Christian Muehl (UT)
Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

2012-24 Laurens van der Werff (UT)
Evaluation of Noisy Transcripts for Spoken Document Retrieval

2012-25 Silja Eckartz (UT)
Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

2012-26 Emile de Maat (UVA)
Making Sense of Legal Text

2012-27 Hayrettin Gurkok (UT)
Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

2012-28 Nancy Pascall (UvT)
Engendering Technology Empowering Women

2012-29 Almer Tigelaar (UT)
Peer-to-Peer Information Retrieval

2012-30 Alina Pommeranz (TUD)
Designing Human-Centered Systems for Reflective Decision Making

2012-31 Emily Bagarukayo (RUN) A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

2012-32 Wietske Visser (TUD)
Qualitative multi-criteria preference representation and reasoning

2012-33 Rory Sie (OUN)
Coalitions in Cooperation Networks (COCOON)

2012-34 Pavol Jancura (RUN)
Evolutionary analysis in PPI networks and applications

2012-35 Evert Haasdijk (VU)
Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

2012-36 Denis Ssebugwawo (RUN)
Analysis and Evaluation of Collaborative Modeling Processes

2012-37 Agnes Nakakawa (RUN)
A Collaboration Process for Enterprise Architecture Creation

2012-38 Selmar Smit (VU)
Parameter Tuning and Scientific Testing in Evolutionary Algorithms

2012-39 Hassan Fatemi (UT)
Risk-aware design of value and coordination networks

2012-40 Agus Gunawan (UvT)
Information Access for SMEs in Indonesia

2012-41 Sebastian Kelle (OU)
Game Design Patterns for Learning

2012-42 Dominique Verpoorten (OU)
Reflection Amplifiers in self-regulated Learning

2012-44 Anna Tordai (VU)
On Combining Alignment Techniques

2012-45 Benedikt Kratz (UvT)
A Model and Language for Business-aware Transactions

2012-46 Simon Carter (UVA)
Exploration and Exploitation of Multilingual Data for Statistical Machine Translation

2012-47 Manos Tsagkias (UVA)
Mining Social Media: Tracking Content and Predicting Behavior

2012-48 Jorn Bakker (TUE)
Handling Abrupt Changes in Evolving Time-series Data

2012-49 Michael Kaisers (UM)
Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

2012-50 Steven van Kervel (TUD)
Ontologogy driven Enterprise Information Systems Engineering

2012-51 Jeroen de Jong (TUD)
Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

====
2013

====

2013-01 Viorel Milea (EUR)
News Analytics for Financial Decision Support

2013-02 Erietta Liarou (CWI)
MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

2013-03 Szymon Klarman (VU)
Reasoning with Contexts in Description Logics

2013-04 Chetan Yadati(TUD)
Coordinating autonomous planning and scheduling

2013-05 Dulce Pumareja (UT)
Groupware Requirements Evolutions Patterns

2013-06 Romulo Goncalves(CWI)
The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

2013-07 Giel van Lankveld (UvT)
Quantifying Individual Player Differences

2013-08 Robbert-Jan Merk(VU)
Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

2013-09 Fabio Gori (RUN)
Metagenomic Data Analysis: Computational Methods and Applications

2013-10 Jeewanie Jayasinghe Arachchige(UvT)
A Unified Modeling Framework for Service Design.

2013-11 Evangelos Pournaras(TUD)
Multi-level Reconfigurable Self-organization in Overlay Services

2013-12 Marian Razavian(VU)
Knowledge-driven Migration to Services

2013-13 Mohammad Safiri(UT)
Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

2013-14 Jafar Tanha (UVA)
Ensemble Approaches to Semi-Supervised Learning Learning

2013-15 Daniel Hennes (UM)
Multiagent Learning - Dynamic Games and Applications

2013-16 Eric Kok (UU)
Exploring the practical benefits of argumentation in multi-agent deliberation

2013-17 Koen Kok (VU)
The PowerMatcher: Smart Coordination for the Smart Electricity Grid

2013-18 Jeroen Janssens (UvT)
Outlier Selection and One-Class Classification

2013-19 Renze Steenhuizen (TUD)
Coordinated Multi-Agent Planning and Scheduling

2013-20 Katja Hofmann (UvA)
Fast and Reliable Online Learning to Rank for Information Retrieval

2013-21 Sander Wubben (UvT)
Text-to-text generation by monolingual machine translation

2013-22 Tom Claassen (RUN)
Causal Discovery and Logic

2013-23 Patricio de Alencar Silva(UvT)
Value Activity Monitoring

2013-24 Haitham Bou Ammar (UM)
Automated Transfer in Reinforcement Learning

2013-25 Agnieszka Anna Latoszek-Berendsen (UM)
Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

2013-26 Alireza Zarghami (UT)
Architectural Support for Dynamic Homecare Service Provisioning

2013-27 Mohammad Huq (UT)
Inference-based Framework Managing Data Provenance

2013-28 Frans van der Sluis (UT)
When Complexity becomes Interesting: An Inquiry into the Information eXperience

2013-29 Iwan de Kok (UT)
Listening Heads

2013-30 Joyce Nakatumba (TUE)
Resource-Aware Business Process Management: Analysis and Support

2013-31 Dinh Khoa Nguyen (UvT)
Blueprint Model and Language for Engineering Cloud Applications

2013-32 Kamakshi Rajagopal (OUN) Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

2013-33 Qi Gao (TUD)
User Modeling and Personalization in the Microblogging Sphere

2013-34 Kien Tjin-Kam-Jet (UT)
Distributed Deep Web Search

2013-35 Abdallah El Ali (UvA)
Minimal Mobile Human Computer Interaction

2013-36 Than Lam Hoang (TUe)
Pattern Mining in Data Streams

2013-37 Dirk Borner (OUN)
Ambient Learning Displays

2013-38 Eelco den Heijer (VU)
Autonomous Evolutionary Art

2013-39 Joop de Jong (TUD)
A Method for Enterprise Ontology based Design of Enterprise Information Systems

2013-40 Pim Nijssen (UM)
Monte-Carlo Tree Search for Multi-Player Games

2013-41 Jochem Liem (UVA)
Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

2013-42 Leon Planken (TUD)
Algorithms for Simple Temporal Reasoning

2013-43 Marc Bron (UVA)
Exploration and Contextualization through Interaction and Concepts

====
2014
====

2014-01 Nicola Barile (UU)
Studies in Learning Monotone Models from Data

2014-02 Fiona Tuliyano (RUN)
Combining System Dynamics with a Domain Modeling Method

2014-03 Sergio Raul Duarte Torres (UT)
Information Retrieval for Children: Search Behavior and Solutions

2014-04 Hanna Jochmann-Mannak (UT)
Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

2014-05 Jurriaan van Reijsen (UU)
Knowledge Perspectives on Advancing Dynamic Capability

2014-06 Damian Tamburri (VU)
Supporting Networked Software Development

2014-07 Arya Adriansyah (TUE)
Aligning Observed and Modeled Behavior

2014-08 Samur Araujo (TUD)
Data Integration over Distributed and Heterogeneous Data Endpoints

2014-09 Philip Jackson (UvT)
Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

2014-10 Ivan Salvador Razo Zapata (VU)
Service Value Networks

2014-11 Janneke van der Zwaan (TUD)
An Empathic Virtual Buddy for Social Support

2014-12 Willem van Willigen (VU)
Look Ma, No Hands: Aspects of Autonomous Vehicle Control

2014-13 Arlette van Wissen (VU)
Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

2014-14 Yangyang Shi (TUD)
Language Models With Meta-information

2014-15 Natalya Mogles (VU)
Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

2014-16 Krystyna Milian (VU)
Supporting trial recruitment and design by automatically interpreting eligibility criteria

2014-17 Kathrin Dentler (VU)
Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

2014-18 Mattijs Ghijsen (VU)
Methods and Models for the Design and Study of Dynamic Agent Organizations

2014-19 Vinicius Ramos (TUE)
Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

2014-20 Mena Habib (UT)
Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

2014-21 Kassidy Clark (TUD)
Negotiation and Monitoring in Open Environments

2014-22 Marieke Peeters (UU)
Personalized Educational Games - Developing agent-supported scenario-based training

2014-23 Eleftherios Sidirourgos (UvA/CWI)
Space Efficient Indexes for the Big Data Era

2014-24 Davide Ceolin (VU)
Trusting Semi-structured Web Data

2014-25 Martijn Lappenschaar (RUN)
New network models for the analysis of disease interaction

2014-26 Tim Baarslag (TUD)
What to Bid and When to Stop

2014-27 Rui Jorge Almeida (EUR)
Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty

2014-28 Anna Chmielowiec (VU)
Decentralized k-Clique Matching

2014-29 Jaap Kabbedijk (UU)
Variability in Multi-Tenant Enterprise Software

2014-30 Peter de Cock (UvT)
Anticipating Criminal Behaviour

2014-31 Leo van Moergestel (UU)
Agent Technology in Agile Multiparallel Manufacturing and Product Support

2014-32 Naser Ayat (UvA)
On Entity Resolution in Probabilistic Data

2014-33 Tesfa Tegegne (RUN)
Service Discovery in eHealth

2014-34 Christina Manteli(VU) The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.

2014-35 Joost van Ooijen (UU)
Cognitive Agents in Virtual Worlds: A Middleware Design Approach

2014-36 Joos Buijs (TUE)
Flexible Evolutionary Algorithms for Mining Structured Process Models

2014-37 Maral Dadvar (UT)
Experts and Machines United Against Cyberbullying

2014-38 Danny Plass-Oude Bos (UT)
Making brain-computer interfaces better: improving usability through post-processing.

2014-39 Jasmina Maric (UvT)
Web Communities, Immigration, and Social Capital

2014-40 Walter Omona (RUN)
A Framework for Knowledge Management Using ICT in Higher Education

2014-41 Frederic Hogenboom (EUR)
Automated Detection of Financial Events in News Text

2014-42 Carsten Eijckhof (CWI/TUD)
Contextual Multidimensional Relevance Models

2014-43 Kevin Vlaanderen (UU)
Supporting Process Improvement using Method Increments

2014-44 Paulien Meesters (UvT)
Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

2014-45 Birgit Schmitz (OUN)
Mobile Games for Learning: A Pattern-Based Approach

2014-46 Ke Tao (TUD)
Social Web Data Analytics: Relevance, Redundancy, Diversity

2014-47 Shangsong Liang (UVA)
Fusion and Diversification in Information Retrieval

====
2015
====

2015-01 Niels Netten (UvA)
Machine Learning for Relevance of Information in Crisis Response

2015-02 Faiza Bukhsh (UvT)
Smart auditing: Innovative Compliance Checking in Customs Controls

2015-03 Twan van Laarhoven (RUN)
Machine learning for network data

2015-04 Howard Spoelstra (OUN)
Collaborations in Open Learning Environments

2015-05 Christoph Bosch(UT)
Cryptographically Enforced Search Pattern Hiding

2015-06 Farideh Heidari (TUD)
Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes

2015-07 Maria-Hendrike Peetz(UvA)
Time-Aware Online Reputation Analysis

2015-08 Jie Jiang (TUD)
Organizational Compliance: An agent-based model for designing and evaluating organizational interactions

2015-09 Randy Klaassen(UT)
HCI Perspectives on Behavior Change Support Systems

2015-10 Henry Hermans (OUN)
OpenU: design of an integrated system to support lifelong learning

2015-11 Yongming Luo(TUE)
Designing algorithms for big graph datasets: A study of computing bisimulation and joins

2015-12 Julie M. Birkholz (VU)
Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks

2015-13 Giuseppe Procaccianti(VU)
Energy-Efficient Software

2015-14 Bart van Straalen (UT)
A cognitive approach to modeling bad news conversations

2015-15 Klaas Andries de Graaf (VU)
Ontology-based Software Architecture Documentation

2015-16 Changyun Wei (UT)
Cognitive Coordination for Cooperative Multi-Robot Teamwork

2015-17 André van Cleeff (UT)
Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

2015-18 Holger Pirk (CWI)
Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories

2015-19 Bernardo Tabuenca (OUN)
Ubiquitous Technology for Lifelong Learners

2015-20 Loís Vanhée(UU)
Using Culture and Values to Support Flexible Coordination

2015-21 Sibren Fetter (OUN)
Using Peer-Support to Expand and Stabilize Online Learning

015-22 Zhemin Zhu(UT)
Co-occurrence Rate Networks

2015-23 Luit Gazendam (VU)
Cataloguer Support in Cultural Heritage

2015-24 Richard Berendsen (UVA)
Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation

2015-25 Steven Woudenberg (UU)
Bayesian Tools for Early Disease Detection

2015-26 Alexander Hogenboom (EUR)
Sentiment Analysis of Text Guided by Semantics and Structure

2015-27 Sándor Héman (CWI)
Updating compressed column-stores

2015-28 Janet Bagorogoza (TiU)
Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

2015-29 Hendrik Baier (UM)
Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

2015-30 Kiavash Bahreini (OUN)
Real-time Multimodal Emotion Recognition in E-Learning

2015-31 Yakup Koç (TUD)
On Robustness of Power Grids

2015-32 Jerome Gard (UL)
Corporate Venture Management in SMEs

2015-33 Frederik Schadd (UM)
Ontology Mapping with Auxiliary Resources

2015-34 Victor de Graaff (UT)
Geosocial Recommender Systems

2015-35 Junchao Xu (TUD)
Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

====
2016
====

2016-01 Syed Saiden Abbas (RUN)
Recognition of Shapes by Humans and Machines

2016-02 Michiel Christiaan Meulendijk (UU)
Optimizing medication reviews through decision support: prescribing a better pill to swallow

2016-03 Maya Sappelli (RUN)
Knowledge Work in Context: User Centered Knowledge Worker Support

2016-04 Laurens Rietveld (VU)
Publishing and Consuming Linked Data

2016-05 Evgeny Sherkhonov (UVA)
Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

2016-06 Michel Wilson (TUD)
Robust scheduling in an uncertain environment

2016-07 Jeroen de Man (VU)
Measuring and modeling negative emotions for virtual training

2016-08 Matje van de Camp (TiU)
A Link to the Past: Constructing Historical Social Networks from Unstructured Data